



Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Adversarial attacks against Windows PE malware detection: A survey of the state-of-the-art

Xiang Ling^a, Lingfei Wu^b, Jiangyu Zhang^d, Zhenqing Qu^d, Wei Deng^d, Xiang Chen^d,
Yaguan Qian^c, Chunming Wu^d, Shouling Ji^d, Tianyue Luo^a, Jingzheng Wu^{a,*}, Yanjun Wu^{a,*}

^a Institute of Software, Chinese Academy of Sciences, Beijing, 100190, Beijing, China

^b Pinterest, New York, 10018, NY, USA

^c Zhejiang University of Science and Technology, Hangzhou, 310023, Zhejiang, China

^d Zhejiang University, Hangzhou, 310027, Zhejiang, China

ARTICLE INFO

Article history:

Received 9 October 2022

Revised 19 December 2022

Accepted 4 February 2023

Available online 7 February 2023

Keywords:

Portable executable

Malware detection

Machine learning

Adversarial machine learning

Deep learning

Adversarial attack

ABSTRACT

Malware has been one of the most damaging threats to computers that span across multiple operating systems and various file formats. To defend against ever-increasing and ever-evolving malware, tremendous efforts have been made to propose a variety of malware detection that attempt to effectively and efficiently detect malware so as to mitigate possible damages as early as possible. Recent studies have shown that, on the one hand, existing machine learning (ML) and deep learning (DL) techniques enable superior solutions in detecting newly emerging and previously unseen malware. However, on the other hand, ML and DL models are inherently vulnerable to adversarial attacks in the form of adversarial examples, which are maliciously generated by slightly and carefully perturbing the legitimate inputs to misbehave. Adversarial attacks are initially studied in the domain of computer vision like image classification, and then quickly extended to other domains, including natural language processing, audio recognition, and even malware detection. In this paper, we focus on malware with the file format of portable executable (PE) in the family of Windows operating systems, namely **Windows PE malware**, as a representative case to study the adversarial attack methods in such adversarial settings. To be specific, we start by first outlining the general learning framework of Windows PE malware detection based on ML/DL and subsequently highlighting three unique challenges of performing adversarial attacks in the context of Windows PE malware. Then, we conduct a comprehensive and systematic review to categorize the state-of-the-art adversarial attacks against PE malware detection, as well as corresponding defenses to increase the robustness of Windows PE malware detection. Finally, we conclude the paper by first presenting other related attacks against Windows PE malware detection beyond the adversarial attacks and then shedding light on future research directions and opportunities.

© 2023 Elsevier Ltd. All rights reserved.

1. Introduction

With the rapid development and advancement of information technology, computer systems are playing an indispensable and ubiquitous role in our daily lives. Meanwhile, the cyberattack that attempts to maliciously exploit the computer system with malicious intentions (e.g., damaging computers or gaining eco-

nomical profits) has been an important type of ever-increasing and constantly evolving security threat in our society. Malware (i.e., short for **Malicious software**) is one of the most common and powerful cyberattacks for attackers to perform malicious activities in computer systems, such as stealing confidential information without permissions, compromising the whole system, and demanding for a large ransom. While malware spans across multiple operating systems (e.g., Windows, Linux, macOS, Android, etc.) with various file formats, such as portable executable (PE), executable and linkable format (ELF), Mach-O, Android application package (APK), and portable document format (PDF), we focus on malware with the PE files in the family of Windows operating systems (namely **Windows PE malware**) in this paper due to the following two reasons. First, malware analysis techniques (e.g., detection methods) for Windows PE files are mostly different

* Corresponding authors.

E-mail addresses: lingxiang@iscas.ac.cn (X. Ling), lwu@email.wm.edu (L. Wu), zhangjiangyu@zju.edu.cn (J. Zhang), quzhenqing@zju.edu.cn (Z. Qu), dengwei@zju.edu.cn (W. Deng), wasdnsxchen@gmail.com (X. Chen), qianyaguan@zust.edu.cn (Y. Qian), wuchunming@zju.edu.cn (C. Wu), sji@zju.edu.cn (S. Ji), tianyue@iscas.ac.cn (T. Luo), jingzheng08@iscas.ac.cn (J. Wu), yanjun@iscas.ac.cn (Y. Wu).

from those for other files like APK and PDF files because their underlying operating systems, the file format, and execution modes are significantly different from each other. Research shows there is no universal malware detection that can satisfactorily detect all kinds of malware, and thus existing literature papers on malware analysis commonly point out what specific operating system they target and what file format they are (Raff and Nicholas, 2020; Ucci et al., 2019). That is the very first and most important reason why we focus on Windows PE malware in this paper. Second, Windows is the most worldwide popular and long-standing operation system for end users while the malware in the file format of PE constitutes the earliest and maximum studied threat in the wild (Schultz et al., 2001). According to the statistics of Kaspersky Lab at the end of 2020, there are an average of 360,000 malware detected by Kaspersky per day and over 90% of which are Windows PE malware (Kaspersky Lab, 2020b). Similar statistical trends have been reported by Kaspersky Lab at the end of 2021 (Kaspersky Lab, 2020a), indicating Windows PE files are still not sufficiently protected until now.

To mitigate and address the ever-increasing number of security threats caused by Windows PE malware, there are tremendous research efforts have been made to detect Windows PE malware effectively and efficiently (Ceschin et al., 2020; Idika and Mathur, 2007; Raff and Nicholas, 2020; Ucci et al., 2019; Ye et al., 2017). In particular, traditional malware detection can be traced back to signature-based malware detection, which determines whether a given suspicious software is malicious or not (*i.e.*, malware or goodware) by comparing its signature with all signatures from the maintained database of malware that has been previously collected and confirmed. It is obvious that the fatal flaw of signature-based malware detection is that it can only detect previously collected and known malware due to the heavy reliance on the malware database. In the last few decades, inspired by the great successes of ML and DL in various long-standing real-world tasks (*e.g.*, computer vision, natural language processing, speech recognition, *etc.*), a variety of ML/DL-based malware detection methods (Ceschin et al., 2020; Raff and Nicholas, 2020; Ucci et al., 2019; Ye et al., 2017) that leverage the high capacity of ML/DL models have been adapted and presented to detect Windows PE malware. In general, these ML/DL-based malware detection methods claim that they can generalize well to predict the new and previously unseen (*i.e.*, zero-day) malware instances due to the inherent generalizability of ML/DL models.

Unfortunately, recent studies have demonstrated that existing ML/DL models are inherently vulnerable to adversarial attacks in the form of adversarial examples, which are maliciously generated by slightly and carefully perturbing the legitimate inputs to confuse the target ML/DL models (Carlini and Wagner, 2017; Goodfellow et al., 2015). Since the creation of adversarial attacks, most research papers focused on studying adversarial attacks in the domain of computer vision (Akhtar and Mian, 2018), *e.g.*, slightly and carefully perturbing a “Persian cat” image x such that the resulting adversarial image x' can be misclassified as a “lionfish” by the target image classifier. Normally, in the context of images, most proposed adversarial attacks resolve to the feature-space attack like various gradient-based methods, which can be directly applied to generate adversarial images. Until now, there have been a large number of adversarial attack methods and corresponding defense mechanisms being proposed by security researchers and practitioners in both academia and industry (Chakraborty et al., 2018; Long et al., 2022; Machado et al., 2021; Serban et al., 2020; Zhang and Li, 2019).

Inspired by those studies of adversarial attacks in the context of images, a natural question arises is that, **is it feasible to perform adversarial attacks against existing malware detection methods, especially against ML/DL based PE malware detection?** To an-

swer the aforementioned question, in recent five years, security researchers and practitioners have proposed lots of adversarial attacks in the context of malware (Al-Dujaili et al., 2018; Anderson et al., 2017; Castro et al., 2019a; 2019b; Ceschin et al., 2019; Chen et al., 2019; 2020a; 2017; Demetrio et al., 2019; 2020a; 2020b; Ebrahimi et al., 2021; Fadadu et al., 2019; Fang et al., 2020; 2019; Fleschman et al., 2018; Hu and Tan, 2017a; 2017b; Kawai et al., 2019; Khormali et al., 2019; Kolosnjaji et al., 2018; Kreuk et al., 2018; Labaca-Castro et al., 2021a; Li and Li, 2021; Li et al., 2020a; Liu et al., 2019; Lucas et al., 2021; Park et al., 2019; Qiao et al., 2022; Rosenberg et al., 2020a; 2020b; 2018; Song et al., 2020; Suciu et al., 2019; Wang and Miikkulainen, 2020; Wu et al., 2018; Yuan et al., 2020; Zhang et al., 2020a; Zhong et al., 2020), requiring that the generated adversarial malware file should not only be misclassified as the “goodware” by the target malware detection model, but also behaves exactly the same as the original malware file. Compared to the adversarial attack in the context of images, exploring the adversarial attack in the context of malware is completely different and extremely challenging due to the highly structured nature of software files like PE files. To put it simply, even though we can employ existing feature-space attacks to generate the “adversarial features of malware”, it is significantly challenging to find the corresponding “adversarial malware file” that can preserve the format, executability, and maliciousness the same as the original malware file.

Related work to this survey. In fact, for the general adversarial attacks and defenses, there are lots of surveys being done on the image, audio, video, text, and graph (Akhtar and Mian, 2018; Alshemali and Kalita, 2020; Chakraborty et al., 2018; Lan et al., 2022; Long et al., 2022; Machado et al., 2021; Serban et al., 2020; Sun et al., 2018; Zeng et al., 2020; Zhang and Li, 2019), but very few surveys focusing on the adversarial attacks in the context of malware (Li et al., 2021; Park and Yener, 2020; Pierazzi et al., 2020). Here, we introduce those closely related surveys and highlight their limitations and differences compared with our survey paper as follows.

- In Pierazzi et al. (2020), Pierazzi *et al.* are the first to present a general mathematical formalization of adversarial attacks in the problem-space and further propose a novel problem-space attack against Android malware detection. Although they identify four key constraints and commonalities among different domains (*i.e.*, image classification, face recognition, code attribution, PDF malware, android malware, *etc.*) in terms of the problem-space adversarial attack, this paper (Pierazzi et al., 2020) is **not a survey paper** as it neither extensively collects all existing research efforts nor systematically categorizes and summarizes these efforts in this research direction.
- In Park and Yener (2020), Park and Yener actually conduct **an incomplete and improper survey** in reviewing existing adversarial attacks against malware classifiers, since the authors mistakenly categorize them into two categories: gradient-driven and problem-driven approaches, which are clearly not sufficient to cover all existing adversarial attacks against malware detection. For instance, the semantics-preserving reinforcement learning (SRL) attack proposed by Zhang et al. (2020a) is neither a gradient-driven nor a problem-driven method. In addition, this paper (Park and Yener, 2020) lacks surveying existing defense methods against such adversarial attacks.
- In Li et al. (2021), Li *et al.* first make a series of formulations and assumptions in the context of adversarial malware detection and then attempt to survey this research field of adversarial malware detection in a broad spectrum of malware formats, including Windows PE, Android Package Kit (APK), and Portable Document Format (PDF), which is supposed to be **too coarse-grained** to fully understand the unique challenges ad-

versarial attacks and defenses for different malware formats. It is well-known that malware detection relies heavily on the specific file formats of malware, and thus existing literature papers on malware analysis commonly point out what specific operating system they target and what file format they are (Raff and Nicholas, 2020; Ucci et al., 2019). Therefore, our paper instead focuses on the malware format of Windows PE, which allows us to specifically identify the distinct characteristics of Windows PE malware and further gain a deeper and thorough understanding of adversarial attacks and defenses in terms of Windows PE malware.

Contributions of this survey. Motivated by the ever-increasing attention of adversarial malware detection, the purpose of this survey is to provide a comprehensive review on the state-of-the-art research efforts of adversarial attacks against Windows PE malware detection as well as corresponding defenses to increase the robustness of existing PE malware detection solutions. We expect that this survey can serve successive researchers and practitioners who are interested in attacking and defending Windows PE malware detection. In addition, this survey also aims to provide the basic principles to solve this challenging question in generating “real” adversarial PE malware rather than unpractical adversarial PE features that violate the principles (i.e., format-preserving, executability-preserving, and maliciousness-preserving). We believe that these principles can constitute a useful guideline when related researchers and practitioners deal with the generic task of malware detection, not only restricted to PE malware detection. To summarize, the key contributions of this survey are as follows.

- To the best of our knowledge, this is the first work that summarizes and highlights three unique challenges of adversarial attacks in the context of Windows PE malware in the wild, i.e., format-preserving, executability-preserving, and maliciousness-preserving.
- We conduct a comprehensive and systematic review for adversarial attacks against Windows PE malware detection and propose a complete taxonomy to categorize the state-of-the-art adversarial attack methods from different viewpoints.
- We summarize the existing adversarial defenses for PE malware detection against the proposed adversarial attacks. In addition, we discuss other types of attacks against Windows PE malware detection beyond adversarial attacks.

Organization. The rest of this paper is organized as follows. We introduce the general layout of PE files and outline the ML/DL-based learning framework of PE malware detection in § 2. In § 3, we manifest three unique challenges of adversarial attacks against PE malware detection compared with the general adversarial attacks in the context of images. § 4 first presents our taxonomy of adversarial attacks against PE malware detection and then gives a detailed introduction to the state of the art. We summarize the existing adversarial defenses for PE malware detection in § 5. In § 6, we first discuss other types of attacks against PE malware detection beyond the adversarial attacks and then point out some research directions and possible opportunities for future work. We conclude our survey paper in § 7.

2. Machine learning and deep learning for PE malware detection

This section aims to provide the basics to understand how to take advantage of ML and DL for malware detection in terms of PE files in the family of Windows operating systems (OSs). In particular, we first introduce the general layout of PE files and PE malware in § 2.1, and then outline the general learning framework of PE malware detection models based on ML/DL in § 2.2.

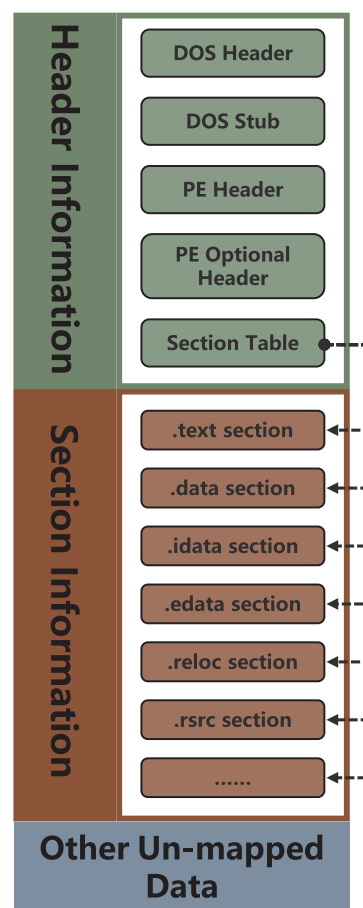


Fig. 1. The general layout of a PE file consists of three groups of information: header information, section information, and the other un-mapped data.

2.1. PE file layout and malware

2.1.1. General layout of PE files

The Portable Executable (PE) format Microsoft, Inc. (2020); Pietrek (2020) is created to provide a common file format for the Microsoft family of Windows OSs to execute code and store essential data that is necessary to execute the code on all supported CPUs, which has made PE the dominant file format of executables on the Windows platform since its creation. As shown in Fig. 1, the general file layout of a PE contains three groups of information, i.e., **header information** (avocado green), **section information** (sandy brown) and the **other un-mapped data** (light blue).

The first group is **header information** which is organized in a fixed manner to tell the OS how to map the PE file into memory for execution. In particular, every PE file starts with the *DOS Header* followed by the *DOS Stub*, both of which are only used for compatibility with the Microsoft disk operating system (MS-DOS) (Paterson, 1983). Following the *DOS Header* and *DOS Stub*, the *PE Header* outlines the overview information about the entire PE file, including the number of sections, the creation time, and various other attributes of PE, etc. Besides that, the subsequent *PE Optional Header* is used to supplementally describe PE files with more than 30 attributes (e.g., size of code, address of entry point, etc.) in more detail. The last component in the header information group usually is the *Section Table*, which provides information about all associated sections, including names, offsets, sizes, and other information. It is worth noting that one PE needs at least one section to be loaded and run.

The second group is **section information** which contains a list of consecutive sections with the executable code and neces-

sary data. In general, the number and the order of these sections are not fixed. Although the name of a section can be customized by the user, Microsoft officially defines several naming conventions based on its semantics and functionalities. In most cases, the “.text” section in PE contains the executable code, while the “.data” section contains necessary data, mainly storing global variables and static variables. The “.idata” and “.edata” sections are used to store the address and size information for the import table and export table, respectively. To be specific, the import table specifies the APIs that will be imported by this executable, while the export table specifies its own functions so that other executables can access them. The “.reloc” section has relocation information to ensure the executable is positioned and executed correctly. The “.rsrc” section contains all resources (e.g., icons, menus, etc.). The last group is the **other un-mapped data** which will not be mapped into memory. In particular, the un-mapped data refers to the chunk of unused bytes, like debug information, at the end of PE files.

Within the family of Windows OSs, PE mainly has two typical and most commonly used file types, i.e., EXEcutable (EXE) and Dynamic Link Library (DLL), which are generally ended with “.exe” and “.dll” as the suffix name. Normally, an “.exe” file can be run independently while a “.dll” file contains the library of functions that other executables can use in the Windows platform (Pietrek, 2020).

2.1.2. PE malware

Malicious software, i.e., malware, is purposely designed and implemented to satisfy the malicious goals and intentions of attackers, e.g., accessing the system without user permissions, stealing private or confidential information, asking for a large ransom, etc. Since the PE file format was first created in the family of Windows OSs, PE files have been widely leveraged by malicious attackers to build PE malware. Until now, according to the security reports from AV-TEST Institute (AV-TEST Institute, 2020) and Avira Protection Labs (Avira, Inc., 2020), PE malware still remains the predominant threat for both personal users and business users in the wild for the following two major reasons. First, the worldwide popularity of the family of Windows OSs and the commonality of PE files inside make the family of Windows OSs, especially the PE file that can be executed, become the main target of attackers for benefit maximization. Second, unlike other file types, PE files can be self-contained, which means that PE malware can include all needed data and does not require additional data to launch the attack. In addition, based on different types of proliferation and different malicious intentions, PE malware can be further briefly classified as viruses, trojans, PUA, worms, adware, ransomware, etc. For more details, we refer interested readers to Souppaya et al. (2013); Ye et al. (2017); Zeidanloo et al. (2010).

2.2. Learning framework for PE malware detection

2.2.1. Overview

To defend against PE malware, i.e., effectively and efficiently detecting PE malware so that potential infections and damages can be mitigated or stopped, there are tremendous research efforts have been made for PE malware detection. Traditional malware detection can actually be traced back to classic signature-based malware detection. To be specific, signature-based malware detection typically maintains a database of signatures of malware that have been previously collected and confirmed. For a given suspicious software like a PE file, signature-based malware detection can determine whether it is malicious or not by comparing its signature with all signatures from the maintained database of malware. Obviously, signature-based malware detection has a fatal drawback in that it heavily relies on the maintained database of malware signatures, so that it can only detect previously collected and known malware.

In recent years, inspired by the great successes of ML and DL techniques in various research fields, various ML/DL-based malware detection methods that leverage the high learning capacity of ML/DL models have been adapted and proposed for PE malware detection. These ML/DL-based malware detection methods normally claim that, as ML/DL models generalize well to predictions of new and unseen instances, they can also generalize to new and previously unseen (i.e., zero-day) malware. Fig. 2 illustrates the overview learning framework of PE malware detection (Ceschin et al., 2020), which generally consists of three steps, including data acquisition, feature engineering as well as learning from models and predictions. In the following, we are going to introduce each step at a glance.

2.2.2. Data acquisition

It is well known that the data quality determines the upper limit of ML/DL models (Cortes et al., 1995). In order to build any malware detection models, it is fundamental to collect and label sufficient PE samples, including both malware and goodware of PE files. However, unlike the benchmark datasets in the field of computer vision or natural language processing (e.g., MNIST (LeCun et al., 1998), CIFAR (Krizhevsky, 2009), ImageNet (Deng et al., 2009), Yelp Review (Inc., 2020), etc.), cybersecurity companies or individuals normally treat PE samples, especially raw files of PE malware, as private property, and thus rarely release them to public. Although a few security institutions or individuals have shared their datasets of PE files (Corvus Forensics, 2021; Ytisf, 2021), there is no established benchmark that has been widely used for most PE malware detection until now. In addition to collecting PE samples, another important process is to distin-

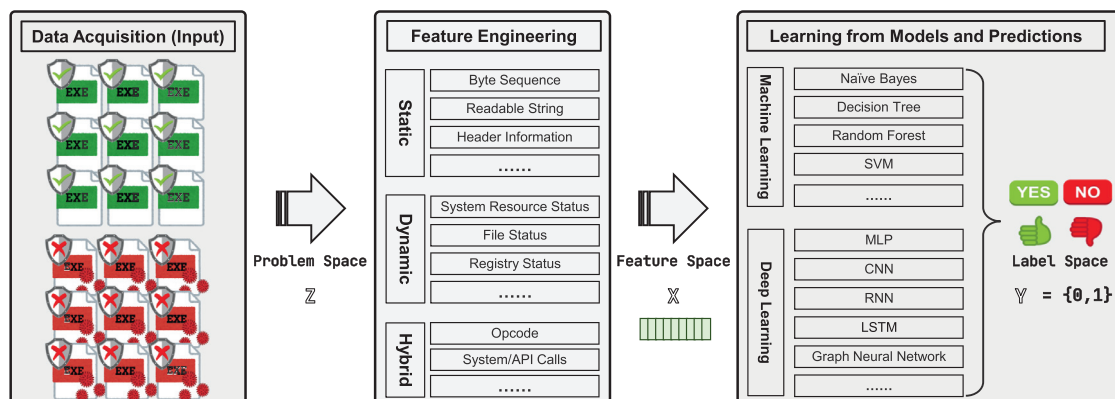


Fig. 2. The Overview Learning Framework of PE Malware Detection.

guish PE malware from all collected PE files, *i.e.*, data labeling. Currently, a common practice to label PE files is to rely on malware analysis services like VirusTotal, which can provide multiple detection results (*i.e.*, whether it is malware) from nearly 70 state-of-the-art anti-virus engines (Chen et al., 2015). Unfortunately, for the same suspicious PE sample, it is well-known that its detection results from different anti-virus engines are somewhat inconsistent with each other. To address these challenges, a variety of methods (Kantchelian et al., 2015; Sebastián et al., 2016; Sebastián and Caballero, 2020; Zhu et al., 2020) have been proposed to unify the detection label by either picking up a “credible” anti-virus (*i.e.*, the most effective and recognized anti-virus software like Kaspersky or Norton) or adopting a voting-based approach from multiple anti-virus engines.

Formally, suppose there is a **problem space** \mathbb{Z} (also known as input space) that contains objects of one specific domain (*e.g.*, PE files, images, texts, *etc.*) from the real-world applications, each object z in the problem space \mathbb{Z} ($z \in \mathbb{Z}$) is associated with a ground-truth label $y \in \mathbb{Y}$, in which \mathbb{Y} denotes the corresponding **label**

space depending on the specific real-world application. In the scenario of PE malware detection, the problem space \mathbb{Z} is referred to the space of all possible PE files, and the label space \mathbb{Y} (*e.g.*, $\mathbb{Y} = \{-1, 1\}$ or $\mathbb{Y} = \{0, 1\}$) denotes the space of detection labels (*e.g.*, -1 or 0 denotes malware and 1 denotes goodware.)

2.2.3. Feature engineering

After collecting and labeling sufficient PE samples, it is necessary and important to perform somewhat feature engineering over all PE samples before inputting them into ML/DL models, as ML/DL models can only accept numeric input. Feature engineering aims to extract the intrinsic properties of PE files that are most likely to be used for distinguishing malware from goodware, and then generates corresponding numeric features for representation. From different perspectives of properties of PE files, there is a large body of work on extracting various features, which can be generally categorized into three broad category: *static* features, *dynamic* features and *hybrid* features (Ceschin et al., 2020; Raff and Nicholas, 2020; Ye et al., 2017) and summarized in Table 1.

Table 1
Common Features and Corresponding Learning Models for PE Malware Detection.

| Features | Category | | Representative ML/DL Model Architectures | Representative PE MalwareDetection Methods* |
|----------------------------------|----------|---------|---|--|
| | Static | Dynamic | | |
| Byte Sequence | ✓ | | Naïve Bayes Murphy et al. (2006), SVM Cortes and Vapnik (1995), DT Witten and Frank (2002), MLP Collobert and Bengio (2004), CNN Krizhevsky et al. (2012), LightGBM Ke et al. (2017). | Schultz et al. (2001), Kolter and Maloof (2006), Saxe and Berlin (2015), Gibert et al. (2018), MalConv Raff et al. (2017), EMBER Anderson and Roth (2018). |
| Readable String | ✓ | | Naïve Bayes Murphy et al. (2006), SVM Cortes and Vapnik (1995), MLP Collobert and Bengio (2004), DT Witten and Frank (2002), RF Ho (1995), LightGBM Ke et al. (2017). | Schultz et al. (2001), SBMDS Ye et al. (2009), Islam et al. (2010), Saxe and Berlin (2015), EMBER Anderson and Roth (2018). |
| Header Information | ✓ | | MLP Collobert and Bengio (2004), LightGBM Ke et al. (2017), Naïve Bayes Murphy et al. (2006), SVM Cortes and Vapnik (1995), DT Witten and Frank (2002). | Saxe and Berlin (2015), EMBER Anderson and Roth (2018), PE-Miner Shafiq et al. (2009). |
| Grayscale Image | ✓ | | kNN Altman (1992), CNN Krizhevsky et al. (2012), SVM Cortes and Vapnik (1995). | Nataraj et al. (2011b), Kim (2017), Visual-AT Liu et al. (2020). |
| CPU/Memory/IO <i>etc.</i> Status | | ✓ | CNN Krizhevsky et al. (2012), kNN Altman (1992), SVM Cortes and Vapnik (1995). | Rieck et al. (2008), AMAL Mohaisen et al. (2015), Abdelsalam et al. (2018). |
| File Status | | ✓ | Hierarchical Clustering Franklin (2005), CNN Krizhevsky et al. (2012), kNN Altman (1992), SVM Cortes and Vapnik (1995), GNN Kipf and Welling (2017). | Bailey et al. (2007), Rieck et al. (2008), AMAL Mohaisen et al. (2015), Abdelsalam et al. (2018), MatchGNet Wang et al. (2019). |
| Registry Status | | ✓ | Hierarchical Clustering Franklin (2005), kNN Altman (1992), SVM Cortes and Vapnik (1995). | Bailey et al. (2007), Rieck et al. (2008), AMAL Mohaisen et al. (2015). |
| Network Status | | ✓ | Hierarchical Clustering Franklin (2005), CNN Krizhevsky et al. (2012), kNN Altman (1992), SVM Cortes and Vapnik (1995), GNN Kipf and Welling (2017). | Bailey et al. (2007), Rieck et al. (2008), AMAL Mohaisen et al. (2015), Abdelsalam et al. (2018), MatchGNet Wang et al. (2019). |
| Opcode | ✓ | ✓ | kNN Altman (1992), SVM Cortes and Vapnik (1995), DT Witten and Frank (2002), RNN Cho et al. (2014), CNN Krizhevsky et al. (2012), Hierarchical Clustering Franklin (2005). | AMCS Ye et al. (2010), Santos et al. (2013), IRMD Zhang et al. (2016), RMVC Sun and Qian (2018). |
| System or API Calls | ✓ | ✓ | RIPPER Cohen (1996), SVM Cortes and Vapnik (1995), Hierarchical Clustering Franklin (2005), CNN Krizhevsky et al. (2012), LSTM Hochreiter and Schmidhuber (1997). | Schultz et al. (2001), SBMDS Ye et al. (2009), Rieck et al. (2008), Qiao et al. (2013), Zhang et al. (2020b). |
| Control Flow Graph | ✓ | ✓ | Naïve Bayes Murphy et al. (2006), SVM Cortes and Vapnik (1995), RF Ho (1995), GNN Kipf and Welling (2017). | Kapoor and Dhavale (2016), MAGIC Yan et al. (2019), MalGraph Ling et al. (2022). |
| Function Call Graph | ✓ | ✓ | RF Ho (1995), AutoEncoder Vincent et al. (2010), CNN Krizhevsky et al. (2012), GNN Kipf and Welling (2017). | Hassen and Chan (2017), DLGraph Jiang et al. (2018), DeepCG Zhao et al. (2019), MalGraph Ling et al. (2022). |

*If the paper does not clearly name the PE malware detection, we use the author name(s) of the paper with its reference.

First of all, static features are directly extracted from the PE samples themselves without actually running them. For instance, byte sequence, readable string, header information, and the grayscale image are commonly used static features for PE malware detection.

- **Byte Sequence:** A PE sample is essentially a binary file, which is typically considered to be a sequence of bytes. Therefore, the byte sequence is the most straightforward and informative way to represent a PE file. In fact, the byte sequence can either be directly input into DL models (Coull and Gardner, 2019; Krčál et al., 2018; Raff et al., 2017), or be further converted into an intermediate representation, e.g., n-grams or entropy of byte sequences (Gibert et al., 2018; Kolter and Maloof, 2006; Saxe and Berlin, 2015; Schultz et al., 2001).
- **Readable String:** A PE file might contain readable strings that reflect its intentions or semantics, like file names, IP addresses, domain names, author signatures, etc. After extracting readable strings in a PE file, their numeric feature representation can be a set of binary attributes (i.e., whether the string exists), frequencies, or even 2D histogram features (Islam et al., 2010; Saxe and Berlin, 2015; Schultz et al., 2001; Ye et al., 2009).
- **Header Information:** As illustrated in Fig. 1, the PE header information occupies an important place to describe and normalize the PE file globally so that it can be executed properly. In particular, simple statistics on PE header information, such as the file size, numbers of sections, sizes of sections, the number of imported or exported functions, etc., are commonly used feature representations for PE malware detection (Anderson and Roth, 2018; Saxe and Berlin, 2015; Shafiq et al., 2009).
- **Grayscale Image:** Since the value range of bytes in a PE file is the same as the pixel value in an image, a visualization-based feature engineering approach is to transform a PE file into a grayscale image, for which each byte in a PE file corresponds to a pixel in an image (Nataraj et al., 2011a). Inspired by the recent great successes of image classification methods, a lot of visualization-based methods have also been proposed for PE malware detection (Kim, 2017; Liu et al., 2020; Nataraj et al., 2011b).

Second, dynamic features refer to those features that can be extracted by first running the executable in an isolated environment (e.g., sandbox, virtual machine, etc.) and then monitoring their runtime status in terms of system resources, files, registries, network, and others.

- **System Resource Status:** The execution of malware inevitably occupies system resources (e.g., CPU, memory, IO, etc.), whose runtime status can be considered as dynamic features for malware detection, as a variety of malware within one specific family might follow a relatively fixed pattern of system resources during execution. In particular, CPU usage, memory usage, and I/O request packets are commonly monitored as dynamic features (Abdelsalam et al., 2018; Mohaisen et al., 2015; Rieck et al., 2008).
- **File Status:** Malware normally needs to operate on files of target users for reaching malicious intentions by attackers. Thus, logging and counting for the files accessed, created, modified, or deleted are commonly used dynamic features in malware detection (Abdelsalam et al., 2018; Bailey et al., 2007; Mohaisen et al., 2015; Rieck et al., 2008; Wang et al., 2019).
- **Registry Status:** Registries that store the system/application-level configurations are important for the family of Windows OSs. The malware could operate on registries with malicious intentions, like self-starting malware. Similar to file status, registry status like counting the registries created, modified, and deleted can also be regarded as dynamic features (Bailey et al., 2007; Mohaisen et al., 2015; Rieck et al., 2008).

- **Network Status:** The spread of malware like trojans and ransomware mainly depends on the network. Taking trojans as an example, they are likely to connect remote servers with certain network ports. Therefore, when diving into the specific aspects of network status, there is a variety of network-level information that can be used for creating a rich set of dynamic features (Abdelsalam et al., 2018; Bailey et al., 2007; Mohaisen et al., 2015; Rieck et al., 2008; Wang et al., 2019), such as the number of distinct IP addresses or certain ports, the number of different HTTP requests (e.g., POST, GET, HEAD, PUT, etc.), the number of common DNS record types (e.g., PTR, CNAMN, SOA, etc.), to name just a few.

Finally, we exemplify four commonly used hybrid features, i.e., opcode, system/API calls, control flow graph (CFG), and function call graph, which can be extracted from executables with either static analysis methods or dynamic analysis methods. For instance, opcodes of executables can be obtained by either extracting from their disassembled instructions or monitoring their runtime instructions in memory.

- **Opcode:** Executables, including malware, can be generally considered as a collection of instructions that are executed in a specific order. In machine assembly language, an instruction consists of an opcode and several operands, in which the opcode specifies the operation to be executed and the operand refers to the corresponding data or its memory location. As prior studies suggest, the opcode distributions of malware statistically differ from goodware, and thus various features are constructed from the opcodes, such as their frequency, n-grams of opcode sequences, or even opcode images (Santos et al., 2013; Sun and Qian, 2018; Ye et al., 2010; Zhang et al., 2016).
- **System/API Calls:** System/API calls refer to how executables interact with system-level or application-level libraries in the family of Windows OSs. Similar to the opcode, various feature representations are thus constructed from system/API calls, such as the frequency of system/API calls, and n-grams of system/API call sequences (Qiao et al., 2013; Rieck et al., 2008; Schultz et al., 2001; Ye et al., 2009; Zhang et al., 2020b).
- **CFG:** The CFG is a graph-based feature representation that is commonly used to characterize the control flow of executables, including PE malware (Kapoor and Dhavale, 2016; Yan et al., 2019). Building from assembly instructions of executables, each node in the CFG represents a sequence of instructions without branching and each edge represents the control flow path between two nodes (Ling et al., 2021a).
- **Function Call Graph:** The function call graph (Ryder, 1979) that attempts to build the caller-callee relation between different functions (including system/API or user-implemented functions), is regarded as a more coarse-grained graph representation compared with CFG (Hassen and Chan, 2017; Jiang et al., 2018; Zhao et al., 2019).

It is worth noting that, on the one hand, we just briefly review and categorize the commonly used features in PE malware detection, and do not attempt to cover all, which is not the goal of our paper. On the other hand, all the features mentioned above are not separate or independent, they are actually mixed for PE malware detection in the wild. In essence, the process of feature engineering can be broadly expressed as a feature mapping function that maps the problem space into the feature space (i.e., the numeric features), which is formulated in Def. 1 as follows.

Def. 1. (Feature Mapping Function) A feature mapping function ϕ is formulated as $\phi: \mathbb{Z} \rightarrow \mathbb{X}$, in which \mathbb{Z} denotes the problem space from a specific real-world application, and \mathbb{X} denotes the corre-

sponding feature space, numerically describing the intrinsic properties of objects in the problem space.

2.2.4. Learning from models and predictions

After extracting and generating the numeric features from the executable, it is necessary to choose a proper ML or DL model for PE malware detection that is generally regarded as a binary classification task, *i.e.*, predicting whether the given executable is malware or goodware. In recent years, with the rapid development and great success of artificial intelligence technology in many fields like computer vision, natural language processing, and even code analysis Ling et al. (2021b), a huge variety of ML/DL models have been continuously proposed, such as Naïve Bayes, SVM, DT, RF, MLP, CNN, RNN, LSTM, or GNN. Regardless of how diverse of feature representation for executables, almost all kinds of ML/DL models mentioned have been used for PE malware detection, as long as the features obtained from feature engineering conform to the input format of the corresponding ML/DL model. For instance, an LSTM model can accept the sequence data as the input, but cannot accept the graph data, while GNN can process the graph data.

In essence, an ML/DL model refers to a mathematical discrimination function f with parameters to map the numeric features of executables into their binary labels (*i.e.*, malware and goodware), which is broadly formulated in Def. 2 as follows.

Def. 2. (Discrimination Function) A discrimination function f can be precisely formulated as $f: \mathbb{X} \rightarrow \mathbb{Y}$, in which \mathbb{X} denotes the feature space and \mathbb{Y} denotes the corresponding label space.

The training process of a malware detection model is to learn the model parameters based on a large number of training samples, so that the malware detection model can approximate the real relationship function between the feature patterns of executables and their binary detection labels. After that, to predict whether a given executable is malware or not, the malware detection model with learned parameters can effectively and efficiently compute the probabilities assigned to both classes of malware and goodware. In order to find the most applicable model, it is actually quite common to test different ML/DL models for PE malware detection depending on the specific task. In Table 1, the last two columns present the representative ML/DL model architectures and corresponding PE malware detection methods with references.

3. Challenges of adversarial attacks for PE malware

In this section, we first introduce the general concept and taxonomy of adversarial attacks that have been originally and extensively studied in the domain of image classification tasks, and then manifest the most unique challenges of adversarial attacks for PE malware when compared with other fields like images, audios, texts, *etc.*

3.1. Adversarial attacks: The general concept and taxonomy

Although recent advances in ML and DL have led to breakthroughs in a variety of long-standing real-world tasks (*e.g.*, computer vision, natural language processing, speech recognition, *etc.*), unfortunately, it has been convincingly demonstrated that existing ML and DL models are inherently vulnerable to adversarial attacks with carefully crafted adversarial examples. In particular, adversarial examples are intentionally and maliciously designed inputs that aim to mislead the given target model in the testing phrase rather than the training phrase. Generally, adversarial attacks can be categorized along multiple different dimensions. In the following part,

we broadly classify adversarial attacks along two dimensions, *i.e.*, adversary's space and adversary's knowledge.

3.1.1. Adversary's space: Feature-space attack versus problem-space attack

Initially, the adversarial example is explored in the context of image classification tasks and is normally addressed in the feature-space domain. In particular, for a given image classification model f and a given input image z with feature representation $x \in \mathbb{X}$ (*i.e.*, $f(x) = y$), the attacker attempts to minimize the distance between x' and x in the feature-space, such that the resulting adversarial example $x' \in \mathbb{X}$ in the feature-space can misclassify the classification model f . This kind of adversarial attack is normally termed as the **feature-space attack**, which is formulated in Eq. (1) as follows.

$$\begin{aligned} \min_{x'} \quad & \text{distance}(x', x) \\ \text{s.t.} \quad & x' \in \mathbb{X} \\ & f(x') = y' \neq y \end{aligned} \quad (1)$$

in which $\text{distance}(x_1, x_2)$ denotes any measurement function of distance between x_1 and x_2 in the feature-space depending on the actual applications. For the feature-space attacks, as the feature representation in the feature-space is normally continuous, most of them generate adversarial examples based on gradient-based methods like FGSM, PGD, C&W, *etc.*

In contrast with the aforementioned feature-space attack, the **problem-space attack** refers to the adversarial attack that is performed in the problem-space, *i.e.*, how to "modify" the real-world input $z \in \mathbb{Z}$ with a minimal cost (*e.g.*, executables, source code, PDF, *etc.*) such that the generated adversarial example $z' \in \mathbb{Z}$ in the problem-space can also misclassify the target model f as follows.

$$\begin{aligned} \min_{z'} \quad & \text{cost}(z', z) \\ \text{s.t.} \quad & z' \in \mathbb{Z} \\ & f(\phi(z')) = y' \neq y \end{aligned} \quad (2)$$

in which $\text{cost}(z_2, z_1)$ denotes any cost function that transforms z_1 into z_2 in the problem-space of the specific application.

When comparing the problem-space attack in Eq. (2) with the feature-space attack in Eq. (1), it is easy to find the most fundamental and noticeable difference between them is, the problem-space attack involves a feature mapping function ϕ that maps the problem-space into the feature-space, which is usually neither invertible nor differentiable. Therefore, the problem-space attack can hardly use gradient-based methods directly to generate adversarial examples. In Fig. 3, we illustrate the connection between the feature-space attack and the problem-space attack.

3.1.2. Adversary's knowledge: White-box attack versus black-box attack

Adversary's knowledge specifies what we assume the adversary knows about the target model to be attacked. In terms of adversary's knowledge, adversarial attacks can be further categorized into the **white-box attack** and the **black-box attack**. To be specific, the white-box attack refers to the scenario that the attackers know all information about the target model (*e.g.*, architectures, weights/parameters, outputs, features, *etc.*) as well as the dataset to train the target model. By contrast, the black-box attack refers to the scenario that the attackers know nothing about the target model except the model output, *e.g.*, the classification label with or without probability.¹ Apparently, the black-box attack is much harder to satisfy than the white-box attack since the white-box attack is equipped with more knowledge about the target model. Besides, there is a wide spectrum between the white-box attack

¹ There is no unified view on whether to treat the scenario of knowing the classification label with probability as the black-box attack.

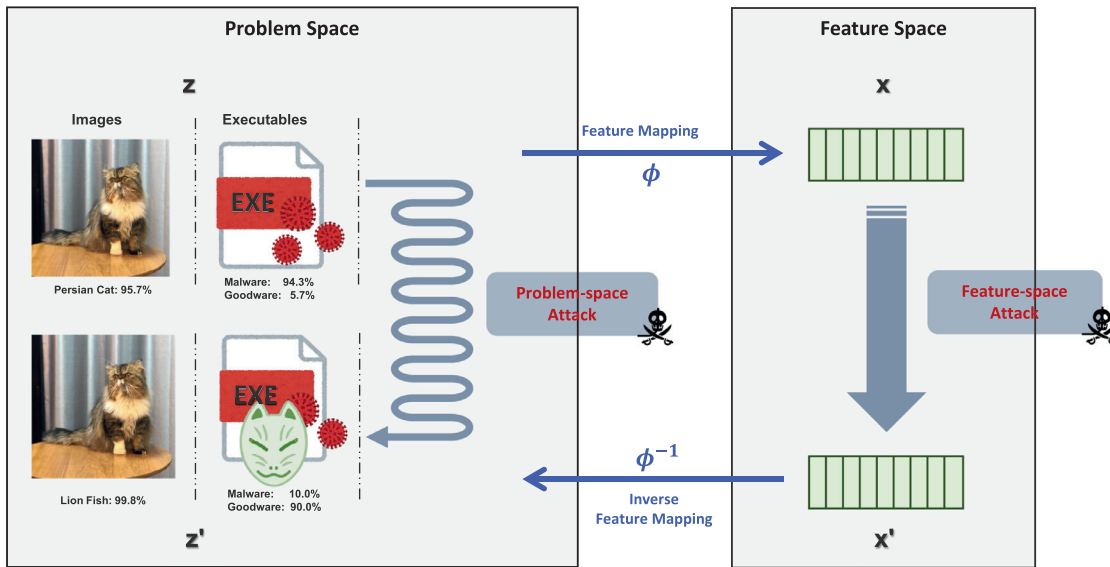


Fig. 3. Illustration of the connection between the **feature-space attack** and the **problem-space attack**, in which the feature mapping function ϕ and the inverse feature mapping function ϕ^{-1} act as bridges for transitions between the feature-space and the problem-space.

and the black-box attack, which is usually broadly referred as the **gray-box** attack. Therefore, for any gray-box attack, it is significantly necessary to show to what extent the adversary knows and does not know about the target model as well as the training dataset.

3.2. Three unique challenges of adversarial attacks for PE malware: From feature-space to problem-space

Originally, adversarial attacks are explored in the domain of image classification tasks and a variety of feature-space attacks are subsequently proposed to generate adversarial examples for the malicious purpose of misclassification, e.g., misclassifying a Persian cat into a lionfish with a high probability of 99.8% as depicted in Fig. 3. Actually, the main reason for the success of directly performing the feature-space attack to generate adversarial examples of images is that, it is easy to find the corresponding image z' from the generate adversarial feature x' via the inverse feature mapping function ϕ^{-1} (i.e., $z' = \phi^{-1}(x')$), as indicated in Fig. 3. However, when considering the adversarial attacks for the PE files, the circumstance becomes completely different and extremely challenging due to the **problem-feature space dilemma** Quiring et al. (2019), which is mainly manifested in the following two aspects.

1. The feature mapping function ϕ_{image} for images is relatively fixed (i.e., an image can be formatted as a two-dimensional array of pixels where each pixel value is a three-dimensional RGB vector with a continuous value between 0 to 255), while the feature mapping function ϕ_{pe} for PE files is not fixed and can take various and diverse approaches of feature engineering as detailed in § 2.2.3. Especially in the setting of black-box attacks, the attacker cannot know the specific feature mapping function ϕ_{pe} for PE files, which greatly increases the difficulty of adversarial attacks for PE files.
2. For images, although the inverse feature mapping function ϕ_{image}^{-1} is not exactly bi-injective (e.g., the pixel value might not be in the range of 0 to 255), it is continuously differentiable, and thus the feature-space attack based on gradients can directly apply on images to generate adversarial examples. How-

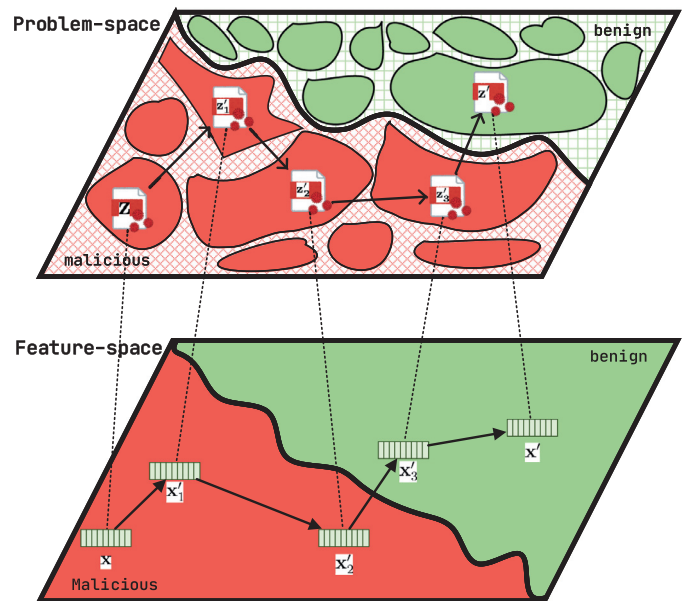


Fig. 4. The schematic illustration of the **feature-space attack** versus the **problem-space attack** for PE malware, in which the original PE malware z is manipulated in the problem-space to continuously generate the adversarial PE malware (i.e., z_1, z_2, z_3 and z'), while the corresponding PE malware feature x in the feature-space is mapped to continuously generate adversarial PE malware features (i.e., x_1, x_2, x_3 and x').

ever, for various different feature mapping functions of PE files, to map a feature vector in the feature-space into an executable in the problem-space, it is almost impossible to find an exact or approximate function of inverse feature mapping ϕ_{pe}^{-1} that is either bi-injective or differentiable.

As depicted in Fig. 4, in order to generate adversarial examples for PE files, although there is a variety of adversarial attacks that exploit the feature-space attacks based on gradients have been proposed, we argue that these adversarial attacks are *impractical* and *unrealistic* against PE malware detection in the wild world.

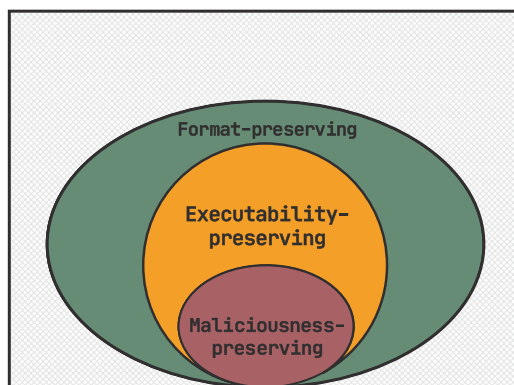


Fig. 5. The schematic illustration of relationship between the three unique challenges of adversarial attacks for PE malware, *i.e.*, **format-preserving**, **executability-preserving** and **maliciousness-preserving**.

This is because what these adversarial attacks generate is the “adversarial PE feature” rather than the “adversarial PE malware” in the end, and an “adversarial PE feature” does not guarantee to correspond to an “adversarial PE malware” due to the following two reasons. On the one hand, it is almost impossible to find a corresponding adversarial PE malware z' based on the generated adversarial PE feature x' , as the inverse feature mapping function ϕ_{pe}^{-1} is normally neither bi-injective nor differentiable. On the other hand, even though we could find the exact PE malware z' in the problem-space that corresponds to the generated adversarial feature x' in the feature-space, there is no guarantee that the found z' is also “adversarial”. Taking the x'_3 in Fig. 4 as an example, although its feature representation x'_3 in feature-space is misclassified as benign (*i.e.*, $f(x'_3) = 0$), but its corresponding PE malware object z'_3 in problem-space is still detected as malicious (*i.e.*, $f(\phi(z'_3)) = 1 \neq 0$).

Therefore, to further generate practical and realistic adversarial PE malware against malware detection in the wild, one of the possible or even the only way so far is to seek for the problem-space attack to generate adversarial PE malware in the problem-space as defined in Eq. (2). To be specific, as depicted in Fig. 4, current problem-space attacks normally attempt to find and apply a series of self-defined problem-space transformations (*i.e.*, T_1 , T_2 , T_3 and T_4) that sequentially transform the original PE malware z into the desired adversarial PE malware z' (*i.e.*, $z \xrightarrow{T_1} z'_1 \xrightarrow{T_2} z'_2 \xrightarrow{T_3} z'_3 \xrightarrow{T_4} z'$), such that \bullet z' is no longer detected as malicious by the target malware detection, and \circ z' maintains the same semantics as the original z . In the following parts, we detail the three unique challenges of maintaining the semantics of adversarial PE malware for practical and realistic adversarial attacks against PE malware detection and present the relationship between the three challenges in Fig. 5.

3.2.1. Challenge 1: Follow the format of PE files (**format-preserving**)

First of all, unlike images, audio, or even texts, PE malware must follow the standard and strict format rules of PE files. As characterized in Fig. 1, the PE file normally has a relatively fixed layout and structure that is necessary to first load the PE file in the system memory and then begin to execute it in the Microsoft family of Windows OSs. Therefore, for PE files, their problem-space transformations should be defined within the scope of the format specification requirements, *i.e.*, format-preserving. For example, one of the transformations that add a new section within the group of Header Information will definitely violate the format of PE files, while adding a new section inside the group of Section Information is acceptable in terms of the format of PE files. In order to overcome the challenge of format-preserving, one of the straightforward and non-trivial approaches is to carefully check and in-

spect each candidate transformation according to the formal specification of PE format under the family of Windows OSs.

3.2.2. Challenge 2: Keep the executability for PE files (**executability-preserving**)

Although one generated adversarial PE malware is format-preserving, it does not necessarily mean it is executability-preserving as well, which is one of the most challenging properties to be addressed in generating adversarial PE malware. This is mainly because, the format of PE files only determines its layout and structure under standard specifications, but cannot particularly determine the concrete content of each element inside the layout and structure, and thus cannot guarantee that they can be properly executed. For example, applying a simple transformation of flipping on byte or even one bit in the “.data” section of a given PE file usually does not violate the format specifications of PE files. However, it is very likely to cause a runtime crash for the transformed PE file as it cannot load the necessary data from the “.data” section, thereby preventing its normal execution.

3.2.3. Challenge 3: Keep the same maliciousness for PE malware (**maliciousness-preserving**)

Recall that the ultimate goal of adversarial attacks against PE malware detection is to generate practical and realistic adversarial PE malware, which could not only misclassify the target PE malware detection model, but also can keep the same maliciousness as the original PE malware. However, again one generated adversarial malware that preserves the executability does not necessarily mean it still preserves the sample maliciousness as the original PE malware, *i.e.*, maliciousness-preserving. Suppose that the generated adversarial malware cannot perform the same maliciousness behaviors (*e.g.*, deleting or encrypting files, modifying registry items, *etc.*) as the original PE malware, it is totally meaningless and unprofitable for the adversary in the wild. Therefore, the property of maliciousness-preserving is another significant challenge to be addressed in generating adversarial PE malware.

In short, to address the two aforementioned challenges of **executability-preserving** and **maliciousness-preserving**, most of the proposed problem-space attacks claim all transformations adopted are format/executability/maliciousness-preserving, but only with limited inspection and analysis on every specific transformation in either concept or empiric. Furthermore, due to the complexity of both PE malware and the corresponding executable environments, we argue that it is almost impossible to theoretically prove whether the proposed adversarial attacks can generate successful adversarial PE malware that satisfies the two properties of executability-preserving and maliciousness-preserving. Alternatively, empirical verification has been reasonably employed in evaluating both executability-preserving and maliciousness-preserving. In particular, if the generated adversarial PE malware can be properly executed and its runtime status is basically consistent with the runtime status of the original PE malware in the same simulated environments (*e.g.*, sandbox), both executability-preserving and maliciousness-preserving can therefore be empirically and reasonably concluded.

4. Adversarial attacks against PE malware detection: The state of the art

In order to explore the most promising advances of adversarial attacks against PE malware detection, in this section, we comprehensively and systematically categorize state-of-the-art adversarial attacks from different viewpoints, *i.e.*, adversary’s knowledge, adversary’s space, target malware detection, and attack strategy. Fig. 6 illustrates the general category of adversarial attacks

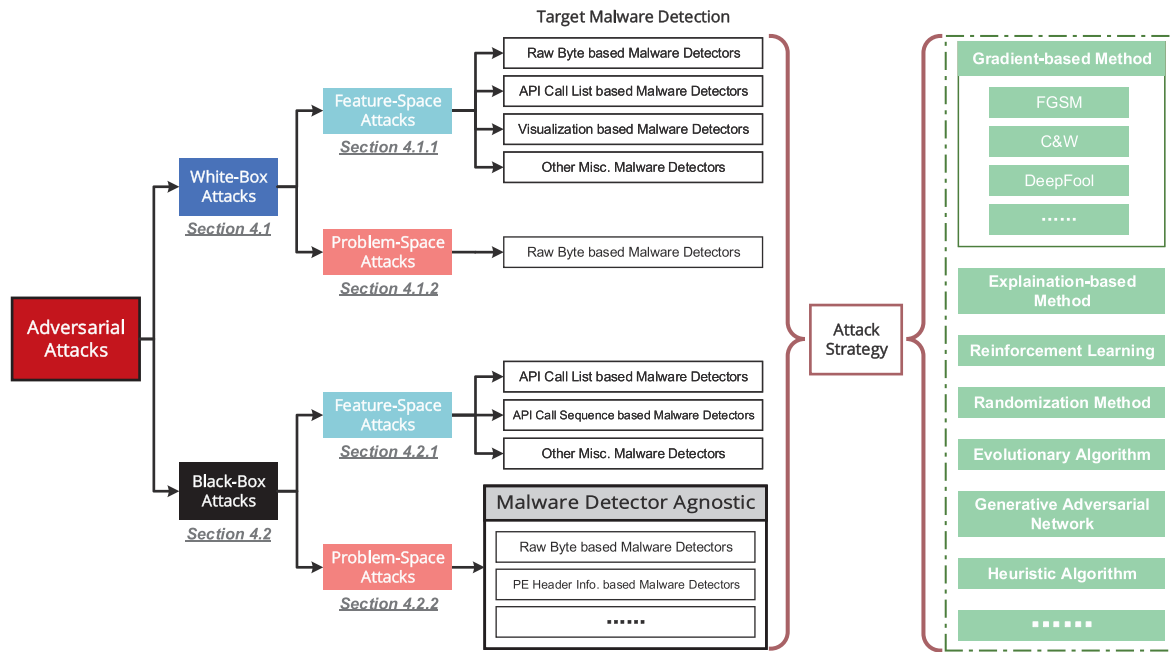


Fig. 6. The General Category of Adversarial Attacks against PE Malware Detection.

against PE malware detection of this paper. In the following subsections, in terms of the adversary’s knowledge (i.e., white-box versus black-box), we will first introduce the white-box adversarial attacks against PE malware detection in § 4.1, and then introduce the black-box adversarial attacks in § 4.2. Finally, we highlight the summary of state-of-the-art adversarial attacks against PE malware detection in § 4.3.

4.1. White-box adversarial attacks against PE malware detection

Recall that in § 3.1, the white-box attack refers to the scenario that the adversary knows full information about the target malware detectors, including architectures, parameters, feature space, the training dataset, etc. In the following parts, we first further divide the white-box attacks into the *feature-space white-box attacks* (in § 4.1.1) and the *problem-space white-box attacks* (in § 4.1.2) based on their corresponding adversary’s space (i.e., feature-space versus problem-space).

4.1.1. Feature-space white-box attacks against PE malware detection

In this part, we focus on the feature-space white-box attacks against PE malware detection, in which all adversarial manipulations (e.g., adding irrelevant API calls) are performed in the feature space of PE malware (e.g., the API call list). In particular, to better understand all feature-space white-box attacks in the face of different kinds of PE malware detection models, we group them into the following categories according to the different types of PE malware detection models, including raw byte based malware detectors, API call list based malware detectors, visualization based malware detectors, and other miscellaneous malware detectors.

Raw Bytes based Malware Detectors. In order to evade the raw bytes based malware detector like MalConv (Raff et al., 2017), Kreuk et al. (2018) consider appending or injecting a sequence of bytes, namely the adversarial payload, to the end of PE malware or the slack region, i.e., existing unused continuous bytes of sections in the middle of PE malware. First, they iteratively generate the adversarial payload with the gradient-based method of FGSM (Goodfellow et al., 2015) in the continuous feature space.

Then, to generate the practical adversarial PE malware, they attempt to reconstruct back to the input problem space by directly searching the closest neighbor to the generated adversarial payload.

Similarly, Suciú et al. (2019) extend the FGSM-based adversarial attacks with two previously proposed strategies (i.e., append-FGSM and slack-FGSM Kreuk et al. (2018)), and further perform a systematic evaluation to compare the effectiveness of both append and slack strategies against MalConv. Their experimental results show that slack-FGSM outperforms append-FGSM with a smaller number of modified bytes. Possible reasons are that the appended bytes of append-FGSM might exceed the maximum size of the model input (e.g., 2MB for MalConv), or that slack-FGSM can make use of surrounding contextual bytes to amplify the power of FGSM since the CNN-based MalConv detector requires the consideration of the contextual bytes within the convolution window.

Chen et al. (2019) suggest that all those adversarial attacks (Kreuk et al., 2018; Suciú et al., 2019) append or inject adversarial bytes that are first initialized by random noises and further iteratively optimized, which might lead to inferior attack performance. To address this issue, Chen et al. propose two novel white-box attacks (i.e., BFA and Enhanced-BFA) with the saliency vector generated by the Grad-CAM approach (Selvaraju et al., 2017). For BFA, it selects the data blocks with significant importance from benign PE files using computed saliency vectors and then appends those data blocks to the end of the original PE malware. Besides that, Enhanced-BFA is presented to use FGSM to iteratively optimize these perturbations generated by BFA. Experimental results show that Enhanced-BFA and BFA have comparative attack performances when the number of appending bytes is large, but Enhanced-BFA is ten times more effective than BFA when the number of appending bytes is small.

Qiao et al. (2022) propose a white-box adversarial attack against raw bytes based malware detectors like MalConv. In particular, it first generates a prototype sample to maximize the output of malware detection model towards the target class (i.e., malware) by directly applying the gradient descent algorithm. Next, to ensure that the generated adversarial PE malware preserves both

executability and maliciousness as the input PE malware, it modifies the modifiable part of input PE malware (*i.e.*, bytes between sections, bytes at the end of PE files, and bytes in the newly added section) in a fine-grained manner under the guidance of the generated prototype sample.

API Call List based Malware Detectors. By flipping the bits of the binary feature vector of malware (“1” denotes the presence of one Windows API call and “0” denotes the absence), Al-Dujaili *et al.* introduce four kinds of white-box adversarial attacks with k multi-steps, namely dFGSM^k, rFGSM^k, BGA^k, and BCA^k Al-Dujaili *et al.* (2018), to attack API call list based malware detectors. To be specific, dFGSM^k and rFGSM^k are two white-box adversarial attacks that are adapted mainly from the FGSM attack in the continuous feature-space (Goodfellow *et al.*, 2015; Kurakin *et al.*, 2017) but extended for the binary feature space via deterministic or randomized rounding, respectively. BGA^k and BCA^k are two gradient ascent based attacks that update multiple bits or one bit in each step, respectively.

As the winner of “Robust Malware Detection Challenge” (MIT-IBM Watson AI Lab, 2019) in both attack and defense tracks, Verwer *et al.* Verwer *et al.* (2020) propose a novel white-box adversarial attack with greedy random accelerated multi-bit search, namely GRAMS, which generates functional adversarial API call features and also builds a more robust malware detector in a standard adversarial training setting. The main idea of GRAMS is to perform a greedy search procedure that explores gradient information as the heuristic to indicate which bits to flip among all the binary search space (*i.e.*, 22,761 API calls). At each iteration, GRAMS flips k bits of API calls that have the largest absolute gradient and exponentially increases or decreases the value of k depending on whether GRAMS finds a better or worse solution. To ensure the functionality of the generated adversarial malware, both Al-Dujaili *et al.* (2018) and Verwer *et al.* (2020) limit the attack to flipping ‘0’ to ‘1’, meaning both of them only add irrelevant API calls.

Visualization based Malware Detectors. Differently, to attack the visualization-based malware detectors, Liu *et al.* (2019) propose the first white-box adversarial attack approach, namely Adversarial Texture Malware Perturbation Attack (ATMPA), based on adversarial attacks in the domain of image classification tasks (Carlini and Wagner, 2017; Goodfellow *et al.*, 2015). In particular, ATMPA first converts the malware sample to a binary texture grayscale image and then manipulates the corresponding adversarial example with subtle perturbations generated from two existing adversarial attack approaches - FGSM (Goodfellow *et al.*, 2015) and C&W (Carlini and Wagner, 2017). However, the major limitation of ATMPA is that the generated adversarial grayscale image of the malware sample destroys the structure of the original malware and thus cannot be executed properly, which makes ATMPA unpractical for real-world PE malware detection.

Similar to ATMPA, Khormali *et al.* present an adversarial attack COPYCAT (Khormali *et al.*, 2019) against visualization based malware detectors with CNNs. COPYCAT also makes use of existing generic adversarial attacks (*e.g.*, FGSM, PGD, C&W, MIM, DeepFool, *etc.*) to generate an adversarial image. After that, COPYCAT appends the generated adversarial image to the end of the original image of malware rather than directly adding it to the original malware image.

Differently, to evade visualization based malware detectors, Park *et al.* (2019) propose another adversarial attack based on the adversarial malware alignment obfuscation (AMAO) algorithm. Specifically, a non-executable adversarial image is first generated by the off-the-shelf adversarial attacks in the field of image classification (Carlini and Wagner, 2017; Goodfellow *et al.*, 2015). Then, in order to attempt to preserve the executability, the adversarial PE malware is finally generated by the AMAO algorithm that mini-

mally inserts semantic NOPs at the available insertion points of the original malware such that the modified PE malware is as similar as possible to the generated no-executable adversarial image.

Other Miscellaneous Malware Detectors. In Li *et al.* (2020a), Li *et al.* first train ML-based malware detection models based on OpCode n-gram features, *i.e.*, the n-gram sequence of operation codes extracted from the disassembled PE file. Then, the authors employ an interpretation model of SHAP (Lundberg and Lee, 2017) to assign each n-gram feature with an importance value and observe that the 4-gram “move + and + or + move” feature is a typical malicious feature as it almost does not appear in the benign PE samples. Thus, based on this observation, the authors consider a generation method of adversarial PE malware by instruction substitution. For instance, the “move + and + or + move” in 10 sampled malware samples can be replaced with “push + pop + and + or + push + pop”, which can be used to bypass the malware detectors in their evaluation.

4.1.2. Problem-space white-box attacks against PE malware detection

Different from these aforementioned feature-space adversarial attacks that operate in the feature space, there is a growing body of work being proposed to perform problem-space adversarial attacks against PE malware detection. To be specific, since it is technically feasible to directly modify the raw bytes of PE files with possible constraints, almost all existing problem-space white-box attacks target at raw byte based malware detectors (*e.g.*, MalConv), which are detailed as follows.

In Kolosnjaji *et al.* (2018), Kolosnjaji *et al.* introduce a gradient-based white-box attack to generate adversarial malware binaries (AMB) against MalConv. To ensure the generated adversarial malware binaries behave identically to the original malware as much as possible, they consider one semantic-preserving manipulation of appending the generated bytes at the end of the original malware file. The appended bytes are generated by a gradient-descent method of optimizing the appended bytes to maximally increase the probability of the appended PE malware that is predicted as goodware.

To unveil the main characteristics learned by MalConv to discriminate PE malware from benign PE files, Demetrio *et al.* (2019) employ the integrated gradient technique Sundararajan *et al.* (2017) for meaningful explanations and find that MalConv is primarily based on the characteristics learned from PE header rather than malicious content in sections. Motivated by the observation, they further present a variant gradient-based white-box attack that is almost the same as Kolosnjaji *et al.* (2018). The only difference is that, AMB (Kolosnjaji *et al.*, 2018) injects adversarial bytes at the end of the PE file while this work is limited to changing the bytes inside the specific DOS header in the PE header.

In Demetrio *et al.* (2020b), Demetrio *et al.* propose a general adversarial attack framework (RAMEn) against PE malware detectors based on two novel functionality-preserving manipulations, namely *Extend* and *Shift*, which inject adversarial payloads by extending the DOS header and shifting the content of the first section in PE files, respectively. In fact, the adversarial payload generation can be optimized in both white-box and black-box settings. For white-box settings, they use the same gradient-based approach as AMB Kolosnjaji *et al.* (2018) to generate adversarial payloads and then inject payloads via *Extend* and *Shift* manipulations. It is simply noted that, for black-box settings, they use the same genetic algorithm as Demetrio *et al.* (2020a) to generate adversarial payloads and then inject them via *Extend* and *Shift* manipulations.

To make it more stealthy than previous adversarial attacks (Demetrio *et al.*, 2019; 2020b; Kolosnjaji *et al.*, 2018), Sharif *et al.* propose a new kind of adversarial attack based on

binary diversification techniques which manipulate the instructions of binaries in a fine-grained function level via two kinds of functionality-preserving transformations, *i.e.*, in-place randomization and code displacement [Lucas et al. \(2021\)](#). In order to guide the transformations that are applied to the PE malware under the white-box setting, they use a gradient ascent optimization to select the transformation only if it shifts its embeddings in a direction similar to the gradient of the attack loss function ([Carlini and Wagner, 2017](#)) with respect to its embeddings.

4.2. Black-box adversarial attacks against PE malware detection

Recall that in § 3.1, the black-box attack refers to the scenario that the adversary knows nothing about the target PE malware detection models except the outputs, *i.e.*, the malicious/benign label with/without probability. In the following parts, we first further divide the black-box attacks into the *feature-space black-box attacks* (in § 4.2.1) and the *problem-space black-box attacks* (in § 4.2.2) based on their corresponding adversary's space (*i.e.*, feature-space versus problem-space).

4.2.1. Feature-space black-box attacks against PE malware detection

In this part, we focus on the feature-space black-box attacks against PE malware detectors, in which all adversarial operations (*e.g.*, insert irrelevant API calls) are performed in the feature space of PE malware (*e.g.*, the API call sequence) instead of being performed in the PE malware themselves. As the feature-space attacks rely on the different feature representations from different types of PE malware detectors, we thus group all existing feature-space black-box attacks into the following categories according to different types of PE malware detectors, including API call list based malware detectors, API call sequence based malware detectors, and other miscellaneous malware detectors.

It is worth noting that we distinguish the API call list from the API call sequence for PE malware detectors as follows. The API call list is a binary feature (*i.e.*, '1' or '0'), indicating whether or not the PE file calls the specific API. The API call sequence represents the sequence of APIs sequentially that is called by the PE file. The API call list can be extracted by either static or dynamic analysis techniques while the API call sequence can only be extracted by dynamic analysis techniques.

API Call List based Malware Detectors. ([Hu and Tan, 2017b](#)) first present a black-box adversarial attack MalGAN based on GAN to attack PE malware detectors based on the API call list. MalGAN assumes the adversary knows the complete feature space (*i.e.*, binary features of the 160 system-level API calls) of the target malware detector and considers only adding some irrelevant API calls into the original malware sample for generating the adversarial malware samples in the feature space. MalGAN first builds a differentiable substitute detector to fit the target black-box malware detector and then trains a generator to minimize the malicious probability of generated adversarial malware predicted by the substitute detector. Subsequently, [Kawai et al. \(2019\)](#) further present an Improved-MalGAN after addressing several issues of MalGAN from a realistic viewpoint. For instance, Improved-MalGAN trains the MalGAN and the target black-box malware detector with different API call lists while the original MalGAN trains with the same API call list.

In [Chen et al. \(2017\)](#), [Chen et al.](#) introduce another black-box adversarial attack, namely EvmAttack. EvmAttack first employs Max-Relevance ([Peng et al., 2005](#)) to calculate the importance of each API call in classifying PE malware or goodware based on the training set and then ranks those API calls into two sets: *M* and *B*. In particular, *M* contains API calls that are highly relevant to malware, while *B* contains API calls that are highly relevant to goodware. Intuitively, EvmAttack is a simple and straightforward attack method

that manipulates the API call list by either adding the API calls in *B* or removing the ones in *M*. Specifically, EvmAttack employs a bidirectional selection algorithm that greedily selects API calls for the manipulation of addition or removal based on the fact that how the manipulation influences the loss of the target PE malware detector.

API Call Sequence based Malware Detectors. Aiming at attacking RNN-based malware detection models that take the API call sequence as the input, [Hu and Tan \(2017a\)](#) propose a generative model based black-box adversarial attack to evade such RNN-based PE malware detectors. In particular, a generative RNN is trained based on PE malware to generate an irrelevant API call sequence that will be inserted into the original API call sequence of the input PE malware, while a substitute RNN model is trained to fit the target RNN-based malware detector based on both benign samples and the gradient information of malware samples from the generative RNN model.

In [Rosenberg et al. \(2018\)](#), [Rosenberg et al.](#) propose a generic end-to-end attack framework, namely GADGET, against state-of-the-art API call sequence-based malware detectors under black-box settings by the transferability property. GADGET is carried out in three steps: i) GADGET first trains a surrogate model to approximate the decision boundaries of the target malware detector by using the Jacobian-based dataset augmentation method ([Papernot et al., 2017](#)). ii) it then performs a white-box attack on the surrogate model to generate the adversarial API call sequence by restricting the insertion of API calls into the original API call sequence. In more detail, GADGET first randomly selects an insert position and then uses a heuristic searching approach to iteratively find and insert the API calls such that the generated adversarial sequence follows the direction indicated by the Jacobian. iii) to generate practical adversarial malware samples from the adversarial API call sequence, GADGET uses a proxy wrapper script to wrap the original malware by calling the additional APIs with valid parameters in the corresponding position based on the generated adversarial API call sequence.

[Fadadu et al. \(2019\)](#) propose an executable level evasion (ELE) attack under black-box settings to evade PE malware detectors based on the API call sequence. The manipulation of ELE is restricted only to the addition of new API calls, which are chosen by maximizing the fraction of sub-sequences that have the added API call in the domain of benign samples and minimizing the fraction of sub-sequences that have the added API call in the domain of malware samples. To further make the modified PE malware can be executed properly, ELE uses a novel IAT (*i.e.*, Import Address Table) hooking method to redirect the control in the adversarial code that is attached to the PE malware. In particular, the adversarial code contains a wrapper function that not only has identical arguments and returns values with the original API function, but also invokes the added API function that is periodically called by the PE malware.

Following GADGET, [Rosenberg et al. \(2020b\)](#) subsequently propose and implement an end-to-end adversarial attack framework, namely BADGER, which consists of a series of query-efficient black-box attacks to misclassify such API call sequence-based malware detector as well as minimize the number of queries. Basically, to preserve the original functionality, the proposed attacks are limited to only inserting API calls with no effect or an irrelevant effect, *e.g.*, opening a non-existent file. To solve the problem of which and where the API calls should be inserted in, the authors propose different attacks with or without knowledge of output probability scores, *i.e.*, the score-based attack and the decision-based attack. For the score-based attack, it uses the self-adaptive uniform mixing evolutionary algorithm ([Dang and Lehre, 2016](#)) to optimize the insertion position with the API calls generated by a pre-trained SeqGAN that has been trained to mimic API call sequences of benign

samples. For the decision-based attack, it selects a random insertion position and then inserts the API call with the same position from the pre-trained SeqGAN. Finally, to make the attacks query-efficient, they first insert a maximum budget of API calls and then employ a logarithmic backtracking method to remove some of the inserted API calls as long as evasion is maintained.

Other Miscellaneous Malware Detectors. (Rosenberg et al., 2020a) present a transferability-based black-box attack against traditional ML-based malware detection models (e.g., EMBER) by modifying the features instead of just adding new features (e.g., adding API calls) like previous attacks. The authors first train a substitute model to fit the target black-box malware detector and then use the explainable algorithm to obtain a list of feature importance of the detection result of the original malware on the substitute model. Subsequently, the authors modify those easily modifiable features with a list of predefined feature values and select a particular value that results in the highest benign probability.

Aiming at attacking graph-based (i.e., CFG) malware detectors Yan et al. (2019), Zhang et al. (2020a) introduce the first semantic-preserving RL-based black-box adversarial attack named SRL. To preserve the original functionality of the malware and retain the structure of the corresponding control flow graph, SRL trains a deep RL agent which could iteratively choose basic blocks in CFG and semantic NOPs for insertion to modify the PE malware until the generated adversarial malware can successfully bypass the target malware detector. Their experimental results show that SRL achieves a nearly 100% attack success rate against two variants of graph neural network based malware detectors.

4.2.2. Problem-space black-box attacks against PE malware detection

In this part, we focus on the problem-space black-box adversarial attacks against PE malware detectors, in which all adversarial operations are performed in the problem space of PE malware under black-box settings, i.e., directly operating the PE malware itself without any consideration of its feature representation (indicating the *problem-space*) as well as the PE malware detectors to be attacked (indicating the *black-box* setting). It means that, in theory, the problem-space black-box adversarial attacks in the scenario of PE malware detection are completely agnostic to specific PE malware detectors. Therefore, regardless of any kind of PE malware detectors, we group the problem-space black-box adversarial attacks according to the attack strategies, including reinforcement learning, randomization, the evolutionary algorithm, GAN, and the heuristic algorithm, which are detailed as follows.

Reinforcement learning based attacks. To expose the weaknesses of current static anti-virus engines, (Anderson et al., 2018; 2017) are the first to study how to automatically manipulate the original PE malware such that the modified PE malware are no longer detected as malicious by the anti-virus engines while do not break the format and functionality. Particularly, with only knowledge of the binary detection output, they propose a completely black-box adversarial attack based on reinforcement learning (RL), namely gym-malware. The gym-malware first defines 10 kinds of format-preserving and functionality-preserving modifications for Windows PE files as the action space available to the agent within the environment. Then, for any given PE malware, gym-malware tries to learn which sequences of modifications in the action space can be used to modify the PE malware, such that the resulting PE malware is most likely to bypass the static anti-virus engines. Although, gym-malware has demonstrated its effectiveness against PE malware detectors, its experimental results also show that RL with an agent of deep Q-network (DQN) or actor-critic with experience replay (ACER) (Sutton and Barto, 2018) offers limited improvement compared with the random policy.

On the basis of gym-malware, there are multiple follow-up work (Chen et al., 2020a; Ebrahimi et al., 2021; Fang et al.,

2020; 2019; Labaca-Castro et al., 2021a; Li and Li, 2021; Wu et al., 2018) proposing problem-space black-box adversarial attacks against static PE malware detection models.

In particular, Wu et al. (2018) propose gym-plus based on gym-malware with the improvement of adding more format-preserving modifications in the action space and their experimental results show that gym-plus with DQN obtains a higher evasion rate than gym-plus with the random policy. Differently, Chen et al. (2020a) propose gym-malware-mini based on gym-malware with a limited and smaller action space. Based on the observation of most of the format-preserving modifications of gym-malware and gym-plus are stochastic in nature (e.g., the appending bytes to the new section are chosen at random for simplicity, etc.) and those modifications are not exactly repeatable, gym-malware-mini makes 6 kinds of random format-preserving modifications to deterministic modifications, making the RL algorithms easier to learn better policies among limited action space.

Besides that, Fang et al. (2019) present a general framework using DQN to evade PE malware detectors, namely DQEAF, which is almost identical to gym-malware in methodology except for three implementation improvements as follows. 1) DQEAF uses a subset of modifications employed in gym-malware and guarantees that all of them would not lead to corruption in the modified malware; 2) DQEAF uses a vector with 513 dimensions as the observed state, which is much lower than that in gym-malware; 3) DQEAF makes priority into consideration during the replay of past transitions. Fang et al. (2020) also observe that the modifications in the action space of gym-malware have some randomness and further found that most effective adversarial malware from gym-malware are generated by UPX pack/unpacked modifications, which could lead to some training problems with RL due to the non-repeatability of those modifications. Thus, they first reduce the action space to 6 categories having certain deterministic parameters and then propose an improved black-box adversarial attack, namely RLAttack-Net, based on the gym-malware implementation.

Ebrahimi et al. (2021) suggest that the RL-based adversarial attacks against PE malware detectors normally employ actor-critic or DQN, which are limited in handling environments with combinatorially large state space. Naturally, they propose an improved RL-based adversarial attack framework of AMG-VAC on the basis of gym-malware Anderson et al. (2018, 2017) by adopting the variational actor-critic, which has been demonstrated to be the state-of-the-art performance in handling environments with combinatorially large state space. As previous RL-based adversarial attacks tend to generate homogeneous and long sequences of transformations, Labaca-Castro et al. (2021a) thus present an RL-based adversarial attack framework of AIMED-RL as well. The main difference between AIMED-RL and other RL-based adversarial attacks is that AIMED-RL introduces a novel penalization to the reward function for increasing the diversity of the generated sequences of transformations while minimizing the corresponding lengths. Li and Li (2021) suggest that existing RL-based adversarial attacks (Anderson et al., 2018; 2017; Fang et al., 2019) employ the artificially defined instant reward function and environment, which are highly subjective and empirical, potentially leading to non-converge of the RL algorithm. Therefore, in order to address the issue of the subjective and empirical reward function, they present an Inverse RL-based adversarial malware generation method, namely AMG-IRL, which could autonomously generate the flexible reward function according to the current optimal strategy.

In short, compared to Chen et al. (2020a); Ebrahimi et al. (2021); Labaca-Castro et al. (2021a); Wu et al. (2018), the adversarial malware samples generated by Fang et al. (2020, 2019); Li and Li (2021) are verified not only for executability within the Cuckoo sandbox (Cuckoo Team, 2020), but also verified for the original maliciousness via comparing the

function call graph between the before and after malware samples with IDA Pro (SHex-Rays, 2020).

Randomization based attacks. To fully automatize the process of generating adversarial malware without corrupting the malware functionality under the black-box setting, Castro et al. (2019b) propose ARMED – automatic random malware modifications to evade detection. ARMED first generates the adversarial PE malware by randomly applying manipulations among 9 kinds of format-preserving modifications from (Anderson et al., 2018; 2017), and then employs the Cuckoo sandbox to test the functionality of the generated adversarial malware samples. In case the functionality test fails, the above steps would be re-start with a new round until the functionality test successes.

Ceschin et al. (2019) find that packing the original PE malware with a distinct packer Cheng et al. (2018); Telock (2020) or embedding the PE malware in a dropper Kwon et al. (2015); Shoair (2020) is an effective approach in bypassing ML-based malware detectors when combined with appending goodware strings to malware. However, some of the generated adversarial PE malware suffer from either not being executed properly or being too large in size. To solve the challenges, the authors implemented a lightweight dropper, namely Dropper, which first creates an entire new PE file to host the original PE malware and then randomly chooses goodware strings to be appended at the end of the newly created PE file.

Similar to the white-box attack of Enhanced-BFA, Chen et al. (2019) also introduce another black-box version of adversarial attack against MalConv. First, it continuously selects data blocks at random from goodware and appends them to PE malware to generate adversarial PE malware. After performing multiple random attacks as above, it then calculates the contribution degree of each data block based on the experience of successful trajectories of data blocks. Finally, it appends the data blocks to the end of PE malware according to the order of their contribution degrees.

In Song et al. (2020), Song et al. propose an automatic black-box attack framework that applies a sequence of actions to rewrite PE malware for evading PE malware detectors. In particular, to generate adversarial malware samples with a minimal set of required actions from macro/micro actions, the authors employ an action sequence minimizer that consists of three steps. First, it randomly selects macro-actions according to the previously updated weights of actions as the action sequence to rewrite the original malware. Second, it tries to remove some unnecessary macro-actions from the action sequence to generate a minimized adversarial malware, and increases the weights of effective actions for updating. Finally, for every macro-action in the minimized adversarial malware, it attempts to replace the macro-action with a corresponding micro-action. Besides that, the proposed framework can also help explain which features are responsible for evasion as every required action in adversarial malware samples corresponds to a type of affected feature.

Evolutionary algorithm based attacks. Following the black-box adversarial attack framework of ARMED Castro et al. (2019b), Castro et al. (2019a) propose AIMED, which employs a genetic programming approach rather than randomization methods to automatically find optimized modifications for generating adversarial PE malware. Firstly, 9 kinds of format-preserving modifications are introduced and applied to the original PE malware to create the population, and then each modified sample in the population is evaluated by a fitness function in terms of functionality, detection ratio, similarity, and the current number of generations. Secondly, if the modified PE malware fails to bypass the PE malware detector, AIMED implements the classic genetic operations, including selection, crossover, and mutation, which are repeated until all generated adversarial PE malware can evade the PE malware detector.

Experiments demonstrate the time of generating the same number of successful adversarial malware is reduced up to 50% compared to the previous random based approach (Castro et al., 2019b).

Similarly, Wang and Miikkulainen (2020) propose to retrain MalConv with the adversarial PE malware, namely MDEA. To generate the adversarial malware samples, MDEA adjusts 10 kinds of format-preserving manipulations from Anderson et al. (2018, 2017) as the action space and employs a genetic algorithm to optimize different action sequences by selecting manipulations from the action space until the generated adversarial malware bypasses the target malware detectors. In particular, as some manipulations are stochastic in nature, MDEA limits each manipulation with a parameter set to make the adversarially trained models converge within an acceptable time. However, the generated adversarial malware samples by MDEA are not tested for functionality like AIMED and ARMED.

To omit the functionality testing (e.g., sandbox required) and further speed up the computation of prior work Castro et al. (2019a,b), Demetrio et al. introduced an efficient black-box attack framework, namely GAMMA (Demetrio et al., 2020a). For GAMMA, the generation approaches of adversarial malware are limited to two types of functionality-preserving manipulations: section injection and padding. Specifically, benign contents are extracted from the goodware as adversarial payloads to inject either into some newly-created sections (section injection) or at the end of the file (padding). The main idea of GAMMA is to formalize the attack as a constrained minimization problem which not only optimizes the probability of evading detection, but also penalizes the size of the injected adversarial payload as a regularization term. To solve the constrained minimization problem, GAMMA also employs a genetic algorithm, including selection, crossover, and mutation, to generate adversarial PE malware to bypass the malware detector with few queries as well as small adversarial payloads.

GAN based attacks. In Yuan et al. (2020), Yuan et al. present GAPGAN, a novel GAN-based black-box adversarial attack framework that is performed at the byte-level against DL-based malware detection, i.e., MalConv. Specifically, GAPGAN first trains a generator and a discriminator concurrently, where the generator intends to generate adversarial payloads that would be appended at the end of original malware samples, and the discriminator attempts to imitate the black-box PE malware detector to recognize both the original PE goodware and the generated adversarial PE malware. After that, GAPGAN uses the trained generator to generate the adversarial payload for every input PE malware and then appends the adversarial payload to the corresponding input PE malware, which generates the adversarial PE malware while preserving its original functionalities. Experiments show that GAPGAN can achieve a 100% attack success rate against MalConv with only appending payloads of 2.5% of the total length of the input malware samples.

Zhong et al. (2020) propose a convolutional generative adversarial network-based (C-GAN) framework, namely MalFox, which can generate functionally indistinguishable adversarial malware samples against realistic black-box antivirus products by perturbing malware files based on packing-related methods. In general, MalFox consists of 5 components: PE Parser, Generator, PE Editor, Detector, and Discriminator, where Detector is the target black-box malware detector (e.g., antivirus product) and Discriminator is an API call-based malware detector, representing the discrimination model in GAN. First, the API call list (i.e., DLL and system functions) of the original malware sample is extracted as a binary feature vector by PE Parser; Second, the Generator takes both the malware feature vector and a sampled 3-dimensional Gaussian noise as input to produce a 3-dimensional perturbation path, indicating whether each of the three perturbation methods (i.e., Obfusmal, Stealmal, and Hollowmal) is adopted. Third, following the produced perturbation path, the PE editor generates the adversarial

malware sample with corresponding perturbation methods. Finally, the Generator will stop training until the generated adversarial PE malware fails to be recognized by Discriminator.

Heuristic based attacks. Aiming to quantify the robustness of PE malware detectors ranging from two ML-based models to four commercial anti-virus engines, [Fleshman et al. \(2018\)](#) propose one novel targeted occlusion black-box attack for comparing with three pre-existing evasion techniques, *i.e.*, random-based gym-malware ([Anderson et al., 2018; 2017](#)), obfuscation through packing ([Aghakhani et al., 2020; Cheng et al., 2018](#)), and malicious ROP injection ([Poulios et al., 2015](#)). For the proposed targeted occlusion attack, it first uses the occlusion binary search method to identify the most important byte region according to the changes of the malicious probability for a given PE malware detector, and then replaces the identified region with completely random bytes or a contiguous byte region selected randomly from benign samples. However, we believe that adversarial malware samples generated by the targeted occlusion attack are destructive because the replacement could prevent the generated malware from being executed, not to mention maintain the original maliciousness.

Based on the two kinds of functionality-preserving transformations (*i.e.*, in-place randomization and code displacement) that manipulate the instructions of binaries in a fine-grained function level, [Lucas et al. \(2021\)](#) also propose another black-box version of adversarial attack based on a general hill-climbing algorithm. This black-box attack is basically similar to the white-box version and the only difference is how the attempted transformation is selected. Specifically, the black-box attack first queries the model after attempting to apply one of the transformations to the PE malware, and then accepts the transformation only if the corresponding benign probability increases.

4.3. Summary of adversarial attacks against PE malware detection

In the research field of PE malware detection, adversarial attack methods have been rapidly proposed and developed in recent years since 2017. For all four adversarial attack categories (*i.e.*, feature-space white-box, problem-space white-box, feature-space black-box, and problem-space black-box) detailed above, their generated adversarial PE malware is becoming more and more practical and effective in attacking the target PE malware detection. We summarize the state-of-the-art adversarial attacks against PE malware detection in [Table 2](#), which demonstrates the adversarial attack methods and their corresponding categories and characteristics in detail.

For all white-box attacks against PE malware detection, regardless of the adversary's space (feature-space or problem-space), it is clearly observed from [Table 2](#) that almost all of the white-box attacks adopt the optimization of gradient-based methods as their attack strategies. Actually, gradient-based optimization and its variants have been widely adopted in the domain of adversarial attacks against image classification models [Carlini and Wagner \(2017\); Goodfellow et al. \(2015\)](#). However, it is infeasible and impractical to directly apply the gradient-based optimization methods to generate "realistic" adversarial PE malware due to the problem-feature space dilemma [Quiring et al. \(2019\)](#). Therefore, it is primarily important to adapt the existing gradient-based methods (*e.g.*, FGSM, C&W) within constraints of feasible transformations (*e.g.*, append adversarial bytes, add irrelevant API calls) according to the different types of PE malware detectors, indicating the white-box attacks normally depend on specific malware detectors.

Compared with the white-box attacks, the black-box attacks are more realistic and practical in the wild due to their minimal reliance on knowledge about the target malware detector. For the

feature-space black-box attacks, as they are actually performed in the feature space of PE files, existing adversarial attack methods generally devise corresponding feasible transformations (*e.g.*, add irrelevant API calls) for PE malware detectors with different feature spaces (*e.g.*, API call list based malware detectors), indicating the black-box attacks are normally malware detector specific. However, for the problem-space black-box attacks with the most strict requirements due to their manipulations in the problem-space, most of them are malware detector agnostic, meaning that these adversarial attack methods can be used to attack any kind of PE malware detectors in theory.

In terms of property preservation (*i.e.*, format, executability, and maliciousness), for all kinds of adversarial attacks against PE malware detection, it is also observed from [Table 2](#) that most of them can only guarantee the format-preserving rather than executability-preserving and maliciousness-preserving. In particular, several adversarial attack methods (*e.g.*, ATMPA [Liu et al. \(2019\)](#), COPYCAT [Khormali et al. \(2019\)](#) and the target occlusion attack ([Fleshman et al., 2018](#))) might destroy the fixed layout and grammar of the PE format that is necessary to load and execute the PE file. On the other hand, for those adversarial attacks like [Castro et al. \(2019a,b\); Ceschin et al. \(2019\); Fang et al. \(2020, 2019\); Li and Li \(2021\); Lucas et al. \(2021\); Song et al. \(2020\)](#), they are verified not only for the executability, but also verified experimentally whether the generated adversarial PE malware keeps the same maliciousness as the original PE malware, which is strongly recommended and advocated in our opinion.

5. Adversarial defenses for PE malware detection

As various adversarial attacks continue to be proposed and evaluated, adversarial defense methods are meanwhile proposed accordingly. In fact, the rapid development of adversarial attacks and counterpart defenses constitutes a constant arms race, meaning a new adversarial attack can easily inspire the defender to devise a novel corresponding defense, and a newly proposed defense method will inevitably lead the attacker to design a new adversarial attack for profit. Therefore, it is important to explore the most promising advances of adversarial defenses for Windows PE malware detection against adversarial attacks. Although there are currently few researchers that specifically and exclusively propose adversarial defenses for Windows PE malware detection, most of the aforementioned research efforts on adversarial attacks might more or less present corresponding defense methods. In this section, we summarize the state-of-the-art adversarial defenses for PE malware detection in recent years, mainly including adversarial training and several other defenses.

5.1. Adversarial training

Adversarial training is one of the mainstream adversarial defenses to resist adversarial attacks regardless of specific applications, *e.g.*, image classification, natural languages processing, speech recognition, *etc.* Intuitively, adversarial training refers to the defense mechanism that attempts to improve the robustness of the ML/DL model by re-training it with the generated adversarial examples with/without original training examples. In the scenario of defending against adversarial attacks for PE malware detection, adversarial training and its variants are also widely adopted and we detail them as follows.

Most studies on adversarial defenses ([Anderson et al., 2018; Chen et al., 2019; 2017; Hu and Tan, 2017b; Wang and Miikkulainen, 2020; Wu et al., 2018; Zhang et al., 2020a](#)) based on adversarial training follow a similar procedure, in which adversarial PE malware or adversarial PE features are first generated and then

Table 2
 Summary of State-of-the-Art Adversarial Attacks against PE Malware Detection. WB/BB is short for the white-box attack and the black-box attack, FS/PS is short for the problem-space attack and the feature-space attack, and ○/● denotes empty/fully preserving the related property. In particular, as it is almost impossible to theoretically prove both properties of executability-preserving and maliciousness-preserving, we thus use ○ to denote that related property is preserved neither conceptually nor empirically, ● to denote it is preserved both conceptually and empirically, and ◐ to denote it is only preserved conceptually but not empirically without experimental verification.

| Attack Names | Year | Adversary's Knowledge | Adversary's Space | PE Malware Detection | | Attack Methods | | Preservation | | |
|---|-------|-----------------------|-------------------|----------------------|---|---|-------------------------------|--------------|---|---|
| | | | | Category | Detection Name | Transformation | Strategy | Format | Executability (with empirical verification) | Maliciousness (with empirical verification) |
| Kreuk et al. (2018) | 2018 | WB | FS | Static | MalConv | Append or inject the adversarial payload | FGSM | ● | ◐ | ◐ |
| Suciu et al. (2019) | 2019 | WB | FS | Static | MalConv | Append or inject adversarial payload | FGSM | ● | ◐ | ◐ |
| BFA, Enhanced-BFA Chen et al. (2019) | 2019 | WB | FS | Static | MalConv | Append the selected or optimized bytes from benign PE files | Grad-CAM or FGSM | ● | ◐ | ◐ |
| Qiao et al. (2022) | 2022 | WB | FS | Static | MalConv | Append or inject adversarial payload | Gradient-based | ● | ◐ | ◐ |
| dFGSM ^k , rFGSM ^k , BGA ^k , BCA ^k Al-Dujaili et al. (2018) | 2018 | WB | FS | Static | API call list based malware detectors | Add irrelevant API calls | Gradient-based | ● | ◐ | ◐ |
| GRAMS Verwer et al. (2020) | 2020 | WB | FS | Static | API call list based malware detectors | Add irrelevant API calls | Gradient-based | ● | ◐ | ◐ |
| ATMPA Liu et al. (2019) | 2019 | WB | FS | Static | Visualization-based malware detectors | Add adversarial noise to the malware image | FGSM, C&W | ○ | ○ | ○ |
| COPYCAT Khormali et al. (2019) | 2019 | WB | FS | Static | Visualization-based malware detectors | Append adversarial noise generated | FGSM, PGD, C&W, MIM, DeepFool | ● | ◐ | ◐ |
| AMAO Park et al. (2019) | 2019 | WB | FS | Static | Visualization-based malware detectors | Insert the semantic NOPs | FGSM, C&W | ● | ◐ | ◐ |
| Li et al. (2020a) | 2020a | WB | FS | Static | Opcode-based malware detectors | Opcode instruction substitution | Interpretation model SHAP | ● | ◐ | ◐ |
| AMB Kolosnjaji et al. (2018) | 2018 | WB | PS | Static | MalConv | Append adversarial bytes | Gradient-based | ● | ◐ | ◐ |
| Demetrio et al. Demetrio et al. (2019) | 2019 | WB | PS | Static | MalConv | Modify specific regions in the PE header | Gradient-based | ● | ◐ | ◐ |
| RAMen Demetrio et al. (2020b) | 2020 | WB | PS | Static | MalConv, Byte-based DNN Model Coull and Gardner (2019) | DOS Header Extension, Content Shifting | Gradient-based | ● | ◐ | ◐ |
| Lucas et al. (2021) | 2021 | WB | PS | Static | MalConv, AvastNet Krčál et al. (2018) | Binary diversification techniques | Gradient-based | ● | ● | ● |
| MalGAN Hu and Tan (2017b) | 2017 | BB w/o prob. | FS | Static | API call list based malware detectors | Add irrelevant API calls | GAN | ● | ◐ | ◐ |
| Improved MalGAN Kawai et al. (2019) | 2019 | BB w/o prob. | FS | Static | API call list based malware detectors | Add irrelevant API calls | GAN | ● | ◐ | ◐ |
| EvnAttack Chen et al. (2017) | 2017 | BB w prob. | FS | Static | API call list based malware detectors | Add or Remove API calls | Greedy Algorithm | ● | ○ | ○ |
| Hu and Tan (2017a) | 2017 | BB w/o prob. | FS | Dynamic | API call sequence based malware detectors | Insert irrelevant API calls | Generative Model | ● | ◐ | ◐ |
| GADGET Rosenberg et al. (2018) | 2018 | BB | FS | Dynamic | API call sequence based malware detectors | Insert irrelevant API calls with IAT Hooking | Transferability, Heuristics | ● | ● | ◐ |
| ELE Fadadu et al. (2019) | 2019 | BB w prob. | FS | Dynamic | API call sequence based malware detectors | Insert API calls with IAT Hooking | Greedy Algorithm | ● | ◐ | ◐ |

(continued on next page)

Table 2 (continued)

| Attack Names | Year | Adversary's Knowledge | Adversary's Space | PE Malware Detection | | Attack Methods | | Preservation | | |
|--|-------|-----------------------|-------------------|----------------------|--|---|---------------------------------|--------------|---|---|
| | | | | Category | Detection Name | Transformation | Strategy | Format | Executability (with empirical verification) | Maliciousness (with empirical verification) |
| BADGER Rosenberg et al. (2020b) Rosenberg et al. (2020a) | 2020 | BB | FS | Dynamic | API call sequence based malware detectors | Insert API calls with IAT Hooking | Evolutionary Algorithm | ● | ⦿ | ⦿ |
| SRL Zhang et al. (2020a) | 2020 | BB w/ prob. | FS | Static | EMBER | Predefined modifiable features | Transferability, Explainable ML | ● | ⦿ | ⦿ |
| gym-malware Anderson et al. (2018, 2017) | 2017 | BB w/o prob. | PS | Static | – | Inject semantic NOPs into CFG blocks | Reinforcement Learning | ● | ⦿ | ⦿ |
| gym-plus Wu et al. (2018) | 2018 | BB w/o prob. | PS | Static | – | format-preserving modifications | Reinforcement Learning | ● | ● | ⦿ |
| gym-malware-mini Chen et al. (2020a) | 2020 | BB w/o prob. | PS | Static | – | format-preserving modifications | Reinforcement Learning | ● | ⦿ | ⦿ |
| DQEA Fang et al. (2019) | 2019 | BB w/o prob. | PS | Static | – | format-preserving modifications | Reinforcement Learning | ● | ● | ● |
| RLAttackNet Fang et al. (2020) | 2020 | BB w/o prob. | PS | Static | – | format-preserving modifications | Reinforcement Learning | ● | ● | ● |
| AMG-VAC Ebrahimi et al. (2021) | 2021 | BB w/o prob. | PS | Static | – | format-preserving modifications | Reinforcement Learning | ● | ⦿ | ⦿ |
| AIMED-RL Labaca-Castro et al. (2021a) | 2021a | BB w/o prob. | PS | Static | – | format-preserving modifications | Reinforcement Learning | ● | ⦿ | ⦿ |
| AMG-IRL Li and Li (2021) | 2021 | BB w/o prob. | PS | Static | – | format-preserving modifications | Reinforcement Learning | ● | ● | ● |
| ARMED Castro et al. (2019b) | 2019 | BB w/ prob. | PS | Static | – | format-preserving modifications | Randomization | ● | ● | ● |
| Dropper Ceschin et al. (2019) | 2019 | BB w/o prob. | PS | Static | – | Append strings from goodwill & Packing | Randomization | ● | ● | ● |
| Chen et al. (2019) | 2019 | BB w/o prob. | PS | Static | MalConv | Append bytes from benign PE files | Experience-based Randomization | ● | ⦿ | ⦿ |
| Song et al. (2020) | 2020 | BB w/o prob. | PS | Static | – | format-preserving (macro & micro) modifications | Weighted Randomization | ● | ● | ● |
| AIMED Castro et al. (2019a) | 2019 | BB w/ prob. | PS | Static | – | format-preserving modifications | Genetic Programming | ● | ● | ● |
| MDEA Wang and Miikkulainen (2020) | 2020 | BB w/ prob. | PS | Static | MalConv | format-preserving modifications | Genetic Algorithm | ● | ⦿ | ⦿ |
| GAMMA Demetrio et al. (2020a) | 2020 | BB w/ prob. | PS | Static | – | Inject and pad sections from benign PE files | Genetic Algorithm | ● | ⦿ | ⦿ |
| GAPGAN Yuan et al. (2020) | 2020 | BB w/o prob. | PS | Static | MalConv | Append bytes to the end | GAN | ● | ⦿ | ⦿ |
| MalFox Zhong et al. (2020) | 2020 | BB w/o prob. | PS | Static | – | Obfuscation-like techniques | Convolutional-GAN | ● | ⦿ | ⦿ |
| Targeted occlusion attack Fleshman et al. (2018) | 2018 | BB w/ prob. | PS | Static | – | Occlusion of important bytes | Binary Search | ○ | ○ | ○ |
| Lucas et al. (2021) | 2021 | BB w/ prob. | PS | Static | MalConv, AvastNet Krčál et al. (2018) | Binary diversification techniques | Hill-climbing algorithm | ● | ● | ● |

re-train a corresponding ML/DL model based on the adversarial PE malware/features with/without the original PE malware/features. For instance, Hu and Tan re-train a random-forest-based malware detector based on the original API calls as well as the adversarial API calls generated by the adversarial attack of MalGAN (Hu and Tan, 2017b). Anderson et al. (2018) first exploit the gym-malware to generate adversarial PE malware, and then re-train the malware detection model of EMBER based on the original PE files and the generated adversarial PE malware. Differently, in addition to adversarial training with adversarial API calls generated by EvnAttack, Chen et al. (2017) also present a new secure-learning framework for PE malware detection, namely SecDefender, which adds a security regularization term by considering the evasion cost of feature manipulations by attackers.

To sum up, for those adversarial defenses based on adversarial training, it is generally observed that, 1) adversarial training is experimentally demonstrated to mitigate one or several adversarial attacks to some extent; 2) adversarial training inevitably introduces significant additional costs in generating adversarial examples during the training process.

5.2. Other defense methods

As the first place in the defender challenge of the Microsoft's 2020 Machine Learning Security Evasion Competition (Microsoft Azure, 2021), Quiring et al. present a combinatorial framework of adversarial defenses – PEberus (Quiring et al., 2020) based on the following three defense methods as follows.

1. A PE file is passed into the semantic gap detectors, which are used to check whether the PE file is maliciously challenged based on three simple heuristics, *i.e.*, slack space, overlay, and duplicate. For instance, a considerably high ratio of the overlay to the overall size usually indicates the PE file might have appended bytes to the overlay.
2. If the PE file is not detected by the semantic gap detectors, PEberus employs the ensemble of existing malware detectors (*e.g.*, EMBER Anderson and Roth (2018), the monotonic skip-gram model (Íncar Romeo et al., 2018), and the signature-based model (VirusTotal, 2020)) and use the max voting for predictions.
3. PEberus also employs a stateful nearest-neighbor detector (Chen et al., 2020b) which continuously checks if the PE file is similar to any of the previously detected malware in the history buffer.

In short, with PEberus, a PE file is predicted as malicious if any of the above three defense methods predicts it as malicious.

Lucas et al. (2021) attempt to mitigate their proposed adversarial attacks by exploiting two defensive mechanisms, *i.e.*, binary normalization and instruction masking. For the defense of binary normalization, it applies the transformation of in-place randomization iteratively into the PE file to reduce its lexicographic representation, so that its potential adversarial manipulations can be undone before inputting it into the downstream PE malware detector. Similarly, the instruction masking defense first selects a random subset of the bytes that pertain to instructions and then masks it with zeros before inputting the PE file into the PE malware detector.

6. Discussions

The previous sections of § 4 and § 5 enable interested readers to have a better and faster understanding with regard to the adversarial attacks and defenses for Windows PE malware detection. In the following subsections, we first present the other related attacks against PE malware detection beyond the aforementioned ad-

versarial attacks in § 6.1 and then shed some light on research directions as well as opportunities for future work in § 6.2.

6.1. Beyond adversarial attacks

6.1.1. Universal adversarial perturbation

Universal Adversarial Perturbation (UAP) is one special type of adversarial attack in which an identical single perturbation can be applied over a large set of inputs for misclassifying the target model in the testing phase. In order to generate the problem-space UAP against PE malware classifiers in the wild, Labaca-Castro et al. (2021b) first prepare a set of available transformations (*e.g.*, adding sections to the PE file, renaming sections, pack/unpacking, *etc.*) for Windows PE files, and then perform a greedy search approach to identify a short sequence of transformations as the UAP for the PE malware classifier. In particular, if the identified short sequence of transformations representing the UAP is applied to any PE malware, then the resulting adversarial PE malware can evade the target PE malware classifier with high probability while preserving the format, executability, and maliciousness.

6.1.2. Training-time poisoning attacks

Different from the adversarial attacks that are performed in the testing phase, the poisoning attacks aim to manipulate the training phase, such that the resulting poisoned model f_b has the same prediction on a clean set of inputs as the cleanly trained model f_c but has an adversarially-chosen prediction on the poisoned input (*i.e.*, the input associated a specific backdoor trigger).

Aiming at misclassifying a specific family of malware as goodware, Sasaki et al. (2019) propose the first poisoning attack against ML-based malware detectors. They assume the strictest attack scenario that the adversary not only has full knowledge of the training dataset and the learning algorithm, but also can manipulate the training samples (*i.e.*, malware) in the feature space as well as their labels. In particular, the poisoning attack framework first selects one specific malware family, and then utilizes the same optimization algorithm like Muñoz-González et al. (2017) to generate the poisoning samples in the feature-space, which are finally adopted to train the poisoned model.

However, poisoning one entire family of malware is much easier to be detected if the defender checks all malware families individually. In contrast, poisoning one specific instance of malware is a more challenging problem, by which any malware associated with a backdoor trigger would be misclassified as goodware, regardless of their malware families. In addition, it is almost impossible to control and manipulate the labeling process for the poisoning data samples for the adversary, since most of the training samples normally are labeled with multiple independent anti-virus engines by security companies and further used to train the malware detection models. Therefore, considering the practicality of the attack scenario in the wild, both following Severi et al. (2021) and Shapira et al. (2020) belong to the clean-label poisoning attack Shafahi et al. (2018), in which the adversary can control and manipulate the poison instance itself, but cannot control the labeling of the poison instance.

In Severi et al. (2021), targeting the feature-based ML malware classifier, Severi et al. consider the scenario where the adversary can know and manipulate the feature-space of software to build the poisoned goodware (*i.e.*, goodware with the backdoor trigger), some of which can be obtained by security companies to train the feature-based ML malware classifier. To be specific, the backdoor trigger is created by presenting a greedy algorithm to select coherent combinations of relevant features and values based on the explainable machine learning technique (*i.e.*, SHAP Lundberg and

Lee (2017)). Experimental results indicate that the case of 1% poison rate and 17 manipulated features results in an attack success rate of about 20%.

Shapira et al. (2020) argue that the attack assumption of feature-space manipulations in Severi et al. (2021) is unrealistic and unreasonable for real-world malware classifiers. In pursuit of a poisoning attack in a problem space, Shapira et al. propose a novel instance poisoning attack by first selecting the goodwill that is most similar to the target malware instance and then adding sections to the goodwill for adversarially training the poisoned model. Actually, the manipulation of adding sections acts as the backdoor trigger, which can remain the functionality of the associated goodwill as well as the malware instance. During the testing phase, the target malware instance associated with the backdoor trigger will be misclassified as benign by the poisoned model.

6.1.3. Model steal attacks

In order to approximate (*i.e.*, steal) a remote deployed detection model f_b of PE malware under a strictly black-box setting, Ali and Eshete Ali and Eshete (2020) propose a best-effort adversarial approximation method, which mainly relies on limited training samples and publicly accessible pre-trained models. The proposed best-effort adversarial approximation method leverages feature representation mapping (*i.e.*, transforming the raw bytes of each PE to an image representation) and cross-domain transferability (*i.e.*, taking advantage of pre-trained models from image classification) to approximate the PE malware detection model f_b by locally training a substitute PE malware detection model f_s . Experimental results show that the approximated model f_s is nearly 90% similar to the black-box model f_b in predicting the same input samples.

6.2. Future directions and opportunities

6.2.1. Strong demands for robust PE malware detection

As described in § 2.2 and § 5, although there are a large number of detection methods for PE malware, there are only very limited adversarial defense methods for building more robust PE malware detection models against adversarial attacks. In general, almost all current adversarial defense methods are empirical defenses based on various heuristics (*e.g.*, adversarial training, input transformation, *etc.*), which are usually only effective for one or a few adversarial attack methods, indicating these empirical defenses are normally attack-specific. On the one hand, with the massive existence and continuous evolution of PE malware and corresponding adversarial malware, we argue that the demand for empirical defense methods against them is also likely to rise accordingly to build more robust PE malware detection models. On the other hand, the substantial work of *robustness certification* applied in image classification tasks suggests a strong demand for PE malware detection models with theoretically guaranteed robustness, as there is no work studying the certified defenses against various adversarial attacks until now. We argue that certifying the robustness of malware detection models not only helps users comprehensively evaluate their effectiveness under any attack, but also increases the transparency of their pricing in cloud services (*e.g.*, malware detection as a service).

6.2.2. Practical and efficient adversarial attacks against commercial anti-viruses in the wild

As introduced and summarized in § 4 and Table 2, the current adversarial attack methods against PE malware detection have been devoted to developing problem-space black-box adversarial attacks, which usually take a similar and typical attack procedure. In general, the attack procedure first defines a set of available

transformations (*e.g.*, inject adversarial payload, insert the semantic NOPs, *etc.*) in the problem-space, and then employs a variety of search strategies (*e.g.*, gradient-based, reinforcement learning, genetic algorithm, *etc.*) to choose a sequence of transformations which can be applied to the original PE malware for generating the adversarial PE malware. Based on the above observation, we argue there is still much room for improving both the effectiveness and efficiency of adversarial attacks against PE malware detection in two aspects: i) devising and defining more practical and stealthier transformations for PE malware. For example, instead of simply inserting the NOPs in the blocks of CFGs (Zhang et al., 2020a) that is easily noticed and removed by defenders, the transformation of splitting one block of CFGs into multiple iteratively called blocks is much stealthier to be noticed and removed. ii) designing and implementing more efficient search strategies to accelerate the generation of adversarial PE malware. We argue that it is quite time-consuming for both RL and genetic algorithm based search strategies.

In addition, it is clearly observed that most existing adversarial attacks target PE malware detection based on static analysis rather than dynamic analysis, which is particularly unknown for both attackers and defenders. However, the mainstream commercial anti-virus software/service used by end users of laptops and servers normally employs a hybrid defense solution with both static analysis and dynamic analysis. Therefore, it is extremely important and urgently demanded to devise and implement practical and efficient adversarial attacks against PE malware detection based on dynamic analysis and commercial anti-viruses.

6.2.3. Lack of benchmark platforms for research

As the continuous emergence of adversarial attacks against Windows PE malware detection has led to the constant arms race between adversarial attacks and defense methods, it is challenging to quantitatively understand the strengths and limitations of these methods due to incomplete and biased evaluation. In fact, in terms of other research fields like adversarial attacks in images, texts, and graphs, there are a large number of corresponding toolboxes and platforms (Li et al., 2020b; Ling et al., 2019; Papernot et al., 2016; Zeng et al., 2020) that have been implemented and open-sourced. Based on these toolboxes and platforms, subsequent practitioners and researchers can not only exploit them to evaluate the actual effectiveness of previously proposed methods, but also take them as a cornerstone to implement their own proposed attacks and defenses, thereby reducing the time consumption for repetitive implementations. Therefore, in order to further advance the research on adversarial attacks against Windows PE malware detection, we argue that it is significantly important to design and implement a benchmark platform with a set of benchmark datasets, representative PE malware detection to be targeted, state-of-the-art adversarial attack & defense methods, performance metrics, and environments for a comprehensive and fair evaluation.

7. Conclusion

In this paper, we conduct a comprehensive review on the state-of-the-art adversarial attacks against Windows PE malware detection as well as the counterpart adversarial defenses. To the best of our knowledge, this is the first work that not only manifests the unique challenges of adversarial attacks in the context of Windows PE malware in the wild, but also systematically categorizes the extensive work from different viewpoints in this research field. Specifically, by comparing the inherent differences between PE malware and images that have been explored originally and traditionally, we present three unique challenges (*i.e.*, format-preserving, executability-preserving, and maliciousness-preserving)

in maintaining the semantics of adversarial PE malware for practical and realistic adversarial attacks. Besides, we review recent research efforts in both adversarial attacks and defenses, and further develop reasonable taxonomy schemes to organize and summarize the existing literature, aiming at making them more concise and understandable for interested readers. Moreover, beyond the aforementioned adversarial attacks, we discuss other types of attacks against Windows PE malware detection and shed some light on future directions. Hopefully, this paper can serve as a useful guideline and give related researchers/practitioners a comprehensive and systematic understanding of the fundamental issues of adversarial attacks against PE malware detection, thereby becoming a starting point to advance this research field.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgement

This project is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No.XDA0320000, the National Natural Science Foundation of China under No.62202457 and No.U1936215, and the project funded by China Postdoctoral Science Foundation under No.2022M713253. Yaguan Qian is also supported by the Key Program of Zhejiang Provincial Natural Science Foundation of China under No.LZ22F020007.

References

Abdelsalam, M., Krishnan, R., Huang, Y., Sandhu, R., 2018. Malware detection in cloud infrastructures using convolutional neural networks. In: International Conference on Cloud Computing. IEEE, San Francisco, CA, USA, pp. 162–169.

Aghakhani, H., Gritti, F., Mecca, F., Lindorfer, M., Ortolani, S., Balzarotti, D., Vigna, G., Kruegel, C., 2020. When malware is packin'heat; limits of machine learning classifiers based on static analysis features. In: Network and Distributed System Security Symposium. The Internet Society, San Diego, California, USA, pp. 1–20.

Akhtar, N., Mian, A., 2018. Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access* 6, 14410–14430.

Al-Dujaili, A., Huang, A., Hemberg, E., O'Reilly, U.-M., 2018. Adversarial deep learning for robust detection of binary encoded malware. In: IEEE Security and Privacy Workshops. IEEE, San Francisco, CA, USA, pp. 76–82.

Ali, A., Eshete, B., 2020. Best-effort adversarial approximation of black-box malware classifiers. In: EAI International Conference on Security and Privacy in Communication Networks. Springer, Washington DC, USA, pp. 318–338.

Alshemali, B., Kalita, J., 2020. Improving the reliability of deep neural networks in nlp: a review. *Knowl Based Syst* 191, 105210.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat* 46 (3), 175–185.

Anderson, H. S., Kharkar, A., Filar, B., Evans, D., Roth, P., 2018. Learning to evade static PE machine learning malware models via reinforcement learning. *ArXiv preprint arXiv:1801.08917*.

Anderson, H.S., Kharkar, A., Filar, B., Roth, P., 2017. Evading machine learning malware detection. In: Black Hat USA. blackhat.com, Las Vegas, NV, USA, pp. 1–6.

Anderson, H. S., Roth, P., 2018. EMBER: an open dataset for training static PE malware machine learning models. *ArXiv preprint arXiv:1804.04637*.

AV-TEST Institute, 2020. Security report 2019/2020. https://www.av-test.org/fileadmin/pdf/security_report/AV-TEST_Security_Report_2019-2020.pdf. Online (last accessed January 15, 2021).

Avira, Inc., 2020. Malware Threat Report: Q4 and 2020 Malware Threat Report. <https://www.avira.com/en/blog/q4-and-2020-malware-threat-report>. Online (last accessed January 17, 2021).

Bailey, M., Oberheide, J., Andersen, J., Mao, Z.M., Jahanian, F., Nazario, J., 2007. Automated classification and analysis of internet malware. In: International Workshop on Recent Advances in Intrusion Detection. Springer, Gold Coast, Australia, pp. 178–197.

Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy. IEEE, San Jose, CA, USA, pp. 39–57.

Castro, R.L., Schmitt, C., Dreo, G., 2019. AIMED: Evolving malware with genetic programming to evade detection. In: International Conference on Trust, Security and Privacy in Computing and Communications / International Conference on Big Data Science and Engineering. IEEE, Rotorua, New Zealand, pp. 240–247.

Castro, R.L., Schmitt, C., Rodosek, G.D., 2019. ARMED: How automatic malware modifications can evade static detection? In: International Conference on Information Management. IEEE, Cambridge, United Kingdom, pp. 20–27.

Ceschin, F., Botacin, M., Gomes, H.M., Oliveira, L.S., Grégio, A., 2019. Shallow security: on the creation of adversarial variants to evade machine learning-based malware detectors. In: Reversing and Offensive-Oriented Trends Symposium. ACM, Vienna, Austria, pp. 1–9.

Ceschin, F., Gomes, H. M., Botacin, M., Bifet, A., Pfahringer, B., Oliveira, L. S., Grégio, A., 2020. Machine learning (in) security: A stream of problems. *ArXiv preprint arXiv:2010.16045*.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D., 2018. Adversarial attacks and defenses: a survey. *ArXiv preprint arXiv:1810.00069*.

Chen, B., Ren, Z., Yu, C., Hussain, I., Liu, J., 2019. Adversarial examples for CNN-based malware detectors. *IEEE Access* 7, 54360–54371.

Chen, J., Jiang, J., Li, R., Dou, Y., 2020. Generating adversarial examples for static PE malware detector based on deep reinforcement learning. *J. Phys. Conf. Ser.* 1575 (1), 012011.

Chen, K., Wang, P., Lee, Y., Wang, X., Zhang, N., Huang, H., Zou, W., Liu, P., 2015. Finding unknown malice in 10 seconds: mass vetting for new threats at the google-play scale. In: USENIX Security Symposium. USENIX Association, Washington, D.C., USA, pp. 659–674.

Chen, L., Ye, Y., Bourlai, T., 2017. Adversarial machine learning in malware detection: Arms race between evasion attack and defense. In: European Intelligence and Security Informatics Conference. IEEE, Athens, Greece, pp. 99–106.

Chen, S., Carlini, N., Wagner, D., 2020. Stateful detection of black-box adversarial attacks. In: ACM Workshop on Security and Privacy on Artificial Intelligence. ACM, Taipei, Taiwan, China, pp. 30–39.

Cheng, B., Ming, J., Fu, J., Peng, G., Chen, T., Zhang, X., Marion, J.-Y., 2018. Towards paving the way for large-scale windows malware analysis: Generic binary unpacking with orders-of-magnitude performance boost. In: ACM SIGSAC Conference on Computer and Communications Security. ACM, Toronto, ON, Canada, pp. 395–411.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: Empirical Methods in Natural Language Processing. ACL, Doha, Qatar, pp. 1724–1734.

Cohen, W.W., 1996. Learning trees and rules with set-valued features. In: AAAI/IAAI. AAAI Press, Portland, Oregon, USA, pp. 709–716.

Collobert, R., Bengio, S., 2004. Links between perceptrons, MLPs and SVMs. In: International Conference on Machine Learning. ACM, Banff, Alberta, Canada, pp. 1–8.

Cortes, C., Jackel, L.D., Chiang, W.-P., et al., 1995. Limits on learning machine accuracy imposed by data quality. In: International Conference on Knowledge Discovery and Data Mining. AAAI Press, Montreal, Canada, pp. 57–62.

Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach Learn* 20 (3), 273–297.

Corvus Forensics, 2021. Virusshare.com - because sharing is caring. <https://virusshare.com/>. Online (last accessed August 25, 2021).

Coull, S.E., Gardner, C., 2019. Activation analysis of a byte-based deep neural network for malware classification. In: IEEE Security and Privacy Workshops. IEEE, San Francisco, CA, UAS, pp. 21–27.

Cuckoo Team, 2020. Cuckoo Sandbox. <https://cuckoosandbox.org>. Online (last accessed September 13, 2020).

Dang, D.-C., Lehre, P.K., 2016. Self-adaptation of mutation rates in non-elitist populations. In: International Conference on Parallel Problem Solving from Nature. Springer, Edinburgh, UK, pp. 803–813.

Demetrio, L., Biggio, B., Lagorio, G., Roli, F., Armando, A., 2019. Explaining vulnerabilities of deep learning to adversarial malware binaries. *ArXiv preprint arXiv:1901.03583*.

Demetrio, L., Biggio, B., Lagorio, G., Roli, F., Armando, A., 2020a. Functionality-preserving black-box optimization of adversarial windows malware. *ArXiv preprint arXiv:2003.13526*.

Demetrio, L., Coull, S. E., Biggio, B., Lagorio, G., Armando, A., Roli, F., 2020b. Adversarial EXamples: A survey and experimental evaluation of practical attacks on machine learning for windows malware detection. *ArXiv preprint arXiv:2008.07125*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Miami, Florida, USA, pp. 248–255.

Ebrahimi, M., Pacheco, J., Li, W., Hu, J.L., Chen, H., 2021. Binary black-box attacks against static malware detectors with reinforcement learning in discrete action spaces. In: IEEE Security and Privacy Workshops. IEEE, Virtual Event, pp. 85–91.

Fadadu, F., Handa, A., Kumar, N., Shukla, S.K., 2019. Evading API call sequence based malware classifiers. In: International Conference on Information and Communications Security. Springer, Beijing, China, pp. 18–33.

Fang, Y., Zeng, Y., Li, B., Liu, L., Zhang, L., 2020. Deepdetectnet vs RLAttacknet: an adversarial method to improve deep learning-based static malware detection model. *PLoS ONE* 15 (4), e0231626.

Fang, Z., Wang, J., Li, B., Wu, S., Zhou, Y., Huang, H., 2019. Evading anti-malware engines with deep reinforcement learning. *IEEE Access* 7, 48867–48879.

Fleshman, W., Raff, E., Zak, R., McLean, M., Nicholas, C., 2018. Static malware detection & subterfuge: Quantifying the robustness of machine learning and current

- anti-virus. In: International Conference on Malicious and Unwanted Software. IEEE, Nantucket, MA, USA, pp. 1–10.
- Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* 27 (2), 83–85.
- Gibert, D., Mateu, C., Planes, J., Vicens, R., 2018. Classification of malware by using structural entropy on convolutional neural networks. In: AAAI Conference on Artificial Intelligence. AAAI Press, New Orleans, Louisiana, USA, pp. 7759–7764.
- Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: International Conference on Learning Representations. OpenReview.net, San Diego, CA, USA, pp. 1–11.
- Hassen, M., Chan, P.K., 2017. Scalable function call graph-based malware classification. In: ACM on Conference on Data and Application Security and Privacy. ACM, Scottsdale, AZ, USA, pp. 239–248.
- Ho, T.K., 1995. Random decision forests. In: International Conference on Document Analysis and Recognition. IEEE, Montreal, Canada, pp. 278–282.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput* 9 (8), 1735–1780.
- Hu, W., Tan, Y., 2017a. Black-box attacks against RNN based malware detection algorithms. ArXiv preprint arXiv:1705.08131.
- Hu, W., Tan, Y., 2017b. Generating adversarial malware examples for black-box attacks based on GAN. ArXiv preprint arXiv:1702.05983.
- Idika, N., Mathur, A.P., 2007. A survey of malware detection techniques. *Purdue University* 48 (2), 32–46.
- Inc., Y., 2020. Yelp open dataset: An all-purpose dataset for learning. <https://www.yelp.com/dataset>. Online (last accessed October 22, 2020).
- İncer Romeo, İ., Theodorides, M., Afroz, S., Wagner, D., 2018. Adversarially robust malware detection using monotonic classification. In: International Workshop on Security and Privacy Analytics. ACM, Tempe, AZ, USA, pp. 54–63.
- Islam, R., Tian, R., Batten, L., Versteeg, S., 2010. Classification of malware based on string and function feature selection. In: Cybercrime and Trustworthy Computing Workshop. IEEE, Ballarat, Australia, pp. 9–17.
- Jiang, H., Turki, T., Wang, J.T., 2018. Dlgraph: Malware detection using deep learning and graph embedding. In: International Conference on Machine Learning and Applications. IEEE, Orlando, FL, USA, pp. 1029–1033.
- Kantchelian, A., Tschantz, M.C., Afroz, S., Miller, B., Shankar, V., Bachwani, R., Joseph, A.D., Tygar, J.D., 2015. Better malware ground truth: Techniques for weighting anti-virus vendor labels. In: ACM Workshop on Artificial Intelligence and Security. ACM, Denver, Colorado, USA, pp. 45–56.
- Kapoor, A., Dhavale, S., 2016. Control flow graph based multiclass malware detection using bi-normal separation. *Def Sci J* 66 (2), 138–145.
- Kaspersky Lab, 2020a. New malicious files discovered daily grow by 5.7% to 380,000 in 2021. https://www.kaspersky.com/about/press-releases/2021_new-malicious-files-discovered-daily-grow-by-57-to-380000-in-2021. Online (last accessed September 14, 2022).
- Kaspersky Lab, 2020b. The number of new malicious files detected every day increases by 5.2% to 360,000 in 2020. https://www.kaspersky.com/about/press-releases/2020_the-number-of-new-malicious-files-detected-every-day-increases-by-52-to-360000-in-2020. Online (last accessed October 1, 2021).
- Kawai, M., Ota, K., Dong, M., 2019. Improved MalGAN: avoiding malware detector by learning cleanware features. In: International Conference on Artificial Intelligence in Information and Communication. IEEE, Okinawa, Japan, pp. 40–45.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., Long Beach, CA, USA, pp. 3146–3154.
- Khormali, A., Abusnaina, A., Chen, S., Nyang, D., Mohaisen, A., 2019. COPYPAT: practical adversarial attacks on visualization-based malware detection. ArXiv preprint arXiv:1909.09735.
- Kim, H.-J., 2017. Image-based malware classification using convolutional neural network. In: Advances in Computer Science and Ubiquitous Computing. Springer, Taichung, Taiwan, China, pp. 1352–1357.
- Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations. OpenReview.net, Toulon, France, pp. 1–14.
- Kolosnjajii, B., Demontis, A., Biggio, B., Maiorca, D., Giacinto, G., Eckert, C., Roli, F., 2018. Adversarial malware binaries: evading deep learning for malware detection in executables. In: European Signal Processing Conference. IEEE, Roma, Italy, pp. 533–537.
- Kolter, J.Z., Maloof, M.A., 2006. Learning to detect and classify malicious executables in the wild. *Journal of Machine Learning Research* 7 (12), 2721–2744.
- Krčál, M., Švec, O., Bálek, M., Jašek, O., 2018. Deep convolutional malware classifiers can learn from raw executables and labels only. In: International Conference on Learning Representations – Workshop Track. OpenReview.net, Vancouver, BC, Canada, pp. 1–4.
- Kreuk, F., Barak, A., Aviv-Reuven, S., Baruch, M., Pinkas, B., Keshet, J., 2018. Deceiving end-to-end deep learning malware detectors using adversarial examples. ArXiv preprint arXiv:1802.04528.
- Krizhevsky, A., 2009. Learning multiple layers of features from tiny images. <http://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., Lake Tahoe, Nevada, United States, pp. 1106–1114.
- Kurakin, A., Goodfellow, I., Bengio, S., 2017. Adversarial examples in the physical world. In: International Conference on Learning Representations. OpenReview.net, Toulon, France, pp. 1–14.
- Kwon, B.J., Mondal, J., Jang, J., Bilge, L., Dumitras, T., 2015. The dropper effect: insights into malware distribution with downloader graph analytics. In: ACM SIGSAC Conference on Computer and Communications Security. ACM, Denver, Colorado, USA, pp. 1118–1129.
- Labaca-Castro, R., Franz, S., Rodosek, G.D., 2021. AIMED-RL: Exploring adversarial malware examples with reinforcement learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Bilbao, Spain, pp. 37–52.
- Labaca-Castro, R., Muñoz-González, L., Pendlebury, F., Rodosek, G. D., Pierazzi, F., Cavallaro, L., 2021b. Universal adversarial perturbations for malware. ArXiv preprint arXiv:2102.06747.
- Lan, J., Zhang, R., Yan, Z., Wang, J., Chen, Y., Hou, R., 2022. Adversarial attacks and defenses in speaker recognition systems: a survey. *J. Syst. Archit.* 127, 102526.
- LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86 (11), 2278–2324.
- Li, D., Li, Q., Ye, Y., Xu, S., 2021. Arms race in adversarial malware detection: a survey. *ACM Computing Surveys (CSUR)* 55 (1), 1–35.
- Li, X., Li, Q., 2021. An IRL-based malware adversarial generation method to evade anti-malware engines. *Computers & Security* 104, 102118.
- Li, X., Qiu, K., Qian, C., Zhao, G., 2020. An adversarial machine learning method based on opcode N-grams feature in malware detection. In: International Conference on Data Science in Cyberspace. IEEE, Hong Kong, China, pp. 380–387.
- Li, Y., Jin, W., Xu, H., Tang, J., 2020b. DeepRobust: A pytorch library for adversarial attacks and defenses. ArXiv preprint arXiv:2005.06149.
- Ling, X., Ji, S., Zou, J., Wang, J., Wu, C., Li, B., Wang, T., 2019. DEEPSEC: A uniform platform for security analysis of deep learning model. In: IEEE Symposium on Security and Privacy. IEEE, San Francisco, USA, pp. 673–690.
- Ling, X., Wu, L., Deng, W., Qu, Z., Zhang, J., Zhang, S., Ma, T., Wang, B., Wu, C., Ji, S., 2022. MalGraph: Hierarchical graph neural networks for robust Windows malware detection. In: IEEE Conference on Computer Communications. IEEE, Virtual Event, pp. 1998–2007.
- Ling, X., Wu, L., Wang, S., Ma, T., Xu, F., Liu, A.X., Wu, C., Ji, S., 2021. Multilevel graph matching networks for deep graph similarity learning. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*.
- Ling, X., Wu, L., Wang, S., Pan, G., Ma, T., Xu, F., Liu, A.X., Wu, C., Ji, S., 2021. Deep graph matching and searching for semantic code retrieval. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15 (5).
- Liu, X., Lin, Y., Li, H., Zhang, J., 2020. A novel method for malware detection on ml-based visualization technique. *Computers & Security* 89, 101682.
- Liu, X., Zhang, J., Lin, Y., Li, H., 2019. ATPPA: Attacking machine learning-based malware visualization detection methods via adversarial examples. In: IEEE/ACM International Symposium on Quality of Service. IEEE, Phoenix, AZ, USA, pp. 1–10.
- Long, T., Gao, Q., Xu, L., Zhou, Z., 2022. A survey on adversarial attacks in computer vision: taxonomy, visualization and future directions. *Computers & Security* 121, 102847.
- Lucas, K., Sharif, M., Bauer, L., Reiter, M.K., Shintre, S., 2021. Malware makeover: Breaking ml-based static analysis by modifying executable bytes. In: ASIA Conference on Computer and Communications Security. ACM, Hongkong, China, pp. 744–758.
- Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. Curran Associates, Inc., Long Beach, CA, USA, pp. 4765–4774.
- Machado, G.R., Silva, E., Goldschmidt, R.R., 2021. Adversarial machine learning in image classification: a survey toward the defenders perspective. *ACM Computing Surveys (CSUR)* 55 (1), 1–38.
- Microsoft Azure, 2021. 2020 machine learning security evasion competition. <https://github.com/Azure/2020-machine-learning-security-evasion-competition>. Online (last accessed January 20, 2021).
- Microsoft, Inc., 2020. PE format. <https://docs.microsoft.com/en-us/windows/win32/debug/pe-format>. Online (last accessed October 22, 2020).
- MIT-IBM Watson AI Lab, 2019. Robust malware detection challenge. 1st Workshop on Adversarial Learning Methods for Machine Learning and Data Mining in KDD 2019 <https://sites.google.com/view/advm1/Home/advm1-2019/advm119-challenge>. Online (last accessed October 15, 2020).
- Mohaisen, A., Alrawi, O., Mohaisen, M., 2015. AMAL: high-fidelity, behavior-based automated malware analysis and classification. *Computers & Security* 52, 251–266.
- Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wonggrassamee, V., Lupu, E.C., Roli, F., 2017. Towards poisoning of deep learning algorithms with back-gradient optimization. In: ACM Workshop on Artificial Intelligence and Security. ACM, Dallas, TX, USA, pp. 27–38.
- Murphy, K.P., et al., 2006. Naive bayes classifiers. Technical Report. University of British Columbia.
- Nataraj, L., Karthikeyan, S., Jacob, G., Manjunath, B.S., 2011. Malware images: visualization and automatic classification. In: International Symposium on Visualization for Cyber Security. ACM, Pittsburgh, PA, USA, pp. 1–7.
- Nataraj, L., Yegneswaran, V., Porras, P., Zhang, J., 2011. A comparative assessment of malware classification using binary texture analysis and dynamic analysis. In: ACM Workshop on Security and Artificial Intelligence. ACM, Chicago, IL, USA, pp. 21–30.
- Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., Xie, C., Sharma, Y., Brown, T., Roy, A., et al., 2016. Technical report on the cleverhans v2.1.0 adversarial examples library. ArXiv preprint arXiv:1610.00768.

- Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A., 2017. Practical black-box attacks against machine learning. In: Asia Conference on Computer and Communications Security. ACM, Abu Dhabi, United Arab Emirates, pp. 506–519.
- Park, D., Khan, H., Yener, B., 2019. Generation & evaluation of adversarial examples for malware obfuscation. In: International Conference on Machine Learning and Applications. IEEE, Boca Raton, FL, USA, pp. 1283–1290.
- Park, D., Yener, B., 2020. A survey on practical adversarial examples for malware classifiers. In: Reversing and Offensive-oriented Trends Symposium. ACM, Vienna, Austria, pp. 23–35.
- Paterson, T., 1983. An inside look at MS-DOS. *Byte* 8 (6), 230.
- Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27 (8), 1226–1238.
- Pierazzi, F., Pendlebury, F., Cortellazzi, J., Cavallaro, L., 2020. Intriguing properties of adversarial ml attacks in the problem space. In: IEEE Symposium on Security and Privacy. IEEE, Virtual Event, pp. 1332–1349.
- Pietrek, M., 2020. Inside Windows: An in-depth look into the Win32 portable executable file format. MSDN Magazine: <https://docs.microsoft.com/en-us/archive/msdn-magazine/2002/february/inside-windows-win32-portable-executable-file-format-in-detail>. Online (last accessed October 22, 2020).
- Poulios, G., Ntantogian, C., Xenakis, C., 2015. ROPinjector: Using return oriented programming for polymorphism and antivirus evasion. In: Black Hat USA. blackhat.com, Las Vegas, NV, USA, pp. 1–11.
- Qiao, Y., Yang, Y., Ji, L., He, J., 2013. Analyzing malware by abstracting the frequent itemsets in api call sequences. In: IEEE International Conference on Trust, Security and Privacy in Computing and Communications. IEEE, Melbourne, Australia, pp. 265–270.
- Qiao, Y., Zhang, W., Tian, Z., Yang, L.T., Liu, Y., Alazab, M., 2022. Adversarial malware sample generation method based on the prototype of deep learning detector. *Computers & Security* 102762.
- Quiring, E., Maier, A., Rieck, K., 2019. Misleading authorship attribution of source code using adversarial learning. In: USENIX Security Symposium. USENIX Association, Santa Clara, CA, USA, pp. 479–496.
- Quiring, E., Pirsch, L., Reimsbach, M., Arp, D., Rieck, K., 2020. Against all odds: Winning the defense challenge in an evasion competition with diversification. ArXiv preprint arXiv:2010.09569.
- Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B., Nicholas, C., 2017. Malware detection by eating a whole EXE. ArXiv preprint arXiv:1710.09435.
- Raff, E., Nicholas, C., 2020. A survey of machine learning methods and challenges for windows malware classification. ArXiv preprint arXiv:2006.09271.
- Rieck, K., Holz, T., Willems, C., Düssel, P., Laskov, P., 2008. Learning and classification of malware behavior. In: International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Springer, Paris, France, pp. 108–125.
- Rosenberg, I., Meir, S., Berrebi, J., Gordon, I., Sicard, G., David, E.O., 2020. Generating end-to-end adversarial examples for malware classifiers using explainability. In: International Joint Conference on Neural Networks. IEEE, Glasgow, United Kingdom, pp. 1–10.
- Rosenberg, I., Shabtai, A., Elovici, Y., Rokach, L., 2020. Query-efficient black-box attack against sequence-based malware classifiers. In: Annual Computer Security Applications Conference. ACM, Virtual Event, pp. 611–626.
- Rosenberg, I., Shabtai, A., Rokach, L., Elovici, Y., 2018. Generic black-box end-to-end attack against state of the art API call based malware classifiers. In: International Symposium on Research in Attacks, Intrusions, and Defenses. Springer, Heraklion, Crete, Greece, pp. 490–510.
- Ryder, B.G., 1979. Constructing the call graph of a program. *IEEE Trans. Software Eng.* 5 (3), 216–226.
- Santos, I., Brezo, F., Ugarte-Pedrero, X., Bringas, P.G., 2013. Opcode sequences as representation of executables for data-mining-based unknown malware detection. *Inf Sci (Ny)* 231, 64–82.
- Sasaki, S., Hidano, S., Uchibayashi, T., Suganuma, T., Hiji, M., Kiyomoto, S., 2019. On embedding backdoor in malware detectors using machine learning. In: International Conference on Privacy, Security and Trust. IEEE, Fredericton, NB, Canada, pp. 1–5.
- Saxe, J., Berlin, K., 2015. Deep neural network based malware detection using two dimensional binary program features. In: International Conference on Malicious and Unwanted Software. IEEE, Fajardo, PR, USA, pp. 11–20.
- Schultz, M.G., Eskin, E., Zadok, F., Stolfo, S.J., 2001. Data mining methods for detection of new malicious executables. In: IEEE Symposium on Security and Privacy. IEEE, Oakland, California, USA, pp. 38–49.
- Sebastián, M., Rivera, R., Kotzias, P., Caballero, J., 2016. AVCLASS: a tool for massive malware labeling. In: International Symposium on Research in Attacks, Intrusions, and Defenses. Springer, Paris, France, pp. 230–253.
- Sebastián, S., Caballero, J., 2020. AVCLASS2: massive malware tag extraction from AV labels. In: Annual Computer Security Applications Conference. ACM, Virtual Event, pp. 42–53.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: International Conference on Computer Vision. IEEE, Venice, Italy, pp. 618–626.
- Serban, A., Poll, E., Visser, J., 2020. Adversarial examples on object recognition: comprehensive survey. *ACM Computing Surveys (CSUR)* 53 (3), 1–38.
- Severi, G., Meyer, J., Coull, S., Oprea, A., 2021. Explanation-guided backdoor poisoning attacks against malware classifiers. In: USENIX Security Symposium. USENIX Association, Virtual Event, pp. 1487–1504. (First submitted to arXiv.org on March 2020 at <https://arxiv.org/abs/2003.01031v1>.)
- Shafahi, A., Huang, W.R., Najibi, M., Suci, O., Studer, C., Dumitras, T., Goldstein, T., 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., Montréal, Canada, pp. 6106–6116.
- Shafiq, M.Z., Tabish, S.M., Mirza, F., Farooq, M., 2009. PE-Miner: Mining structural information to detect malicious executables in realtime. In: International Workshop on Recent Advances in Intrusion Detection. Springer, Saint-Malo, France, pp. 121–141.
- Shapira, T., Berend, D., Rosenberg, I., Liu, Y., Shabtai, A., Elovici, Y., 2020. Being single has benefits. Instance poisoning to deceive malware classifiers. ArXiv preprint arXiv:2010.16323.
- SHEX-Rays, 2020. IDA Pro. <https://www.hex-rays.com/products/ida/>. Online (last accessed September 13, 2020).
- Shoair, K., 2020. DrOp1t-framework. <https://github.com/D4Vinci/DrOp1t-Framework>. Online (last accessed October 25, 2020).
- Song, W., Li, X., Afroz, S., Garg, D., Kuznetsov, D., Yin, H., 2020. Automatic generation of adversarial examples for interpreting malware classifiers. ArXiv preprint arXiv:2003.03100.
- Souppaya, M., Scarfone, K., et al., 2013. Guide to malware incident prevention and handling for desktops and laptops. NIST Special Publication 800, 83.
- Suci, O., Coull, S.E., Johns, J., 2019. Exploring adversarial examples in malware detection. In: IEEE Security and Privacy Workshops. IEEE, San Francisco, CA, USA, pp. 8–14.
- Sun, G., Qian, Q., 2018. Deep learning and visualization for identifying malware families. *IEEE Trans Dependable Secure Comput* 18, 283–295.
- Sun, L., Dou, Y., Yang, C., Wang, J., Yu, P. S., He, L., Li, B., 2018. Adversarial attack and defense on graph data: A survey. ArXiv preprint arXiv:1812.10528.
- Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: International Conference on Machine Learning. PMLR, Sydney, NSW, Australia, pp. 3319–3328.
- Sutton, R.S., Barto, A.G., 2018. Reinforcement learning: An introduction. MIT Press, Cambridge, Massachusetts, USA / London, England, Second edition (in progress)
- Telock, 2020. Telock Version 0.98 for Windows. <https://www.softpedia.com/get/Programming/Packers-Crypters-Protectors/Telock.shtml>. Online (last accessed October 25, 2020).
- Ucci, D., Aniello, L., Baldoni, R., 2019. Survey of machine learning techniques for malware analysis. *Computers & Security* 81, 123–147.
- Verwer, S., Nadeem, A., Hammerschmidt, C., Bliet, L., Al-Dujaili, A., O'Reilly, U.-M., 2020. The robust malware detection challenge and greedy random accelerated multi-bit search. In: ACM Workshop on Artificial Intelligence and Security (AISec). ACM, Virtual Event, pp. 61–70.
- Vincent, P., Laroche, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., Bottou, L., 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11 (12), 3371–3408.
- VirusTotal, 2020. YARA in a nutshell. <https://github.com/virustotal/yara>. Online (last accessed December 15, 2020).
- Wang, S., Chen, Z., Yu, X., Li, D., Ni, J., Tang, L., Gui, J., Li, Z., Chen, H., Yu, P.S., 2019. Heterogeneous graph matching networks for unknown malware detection. In: International Joint Conference on Artificial Intelligence. ijcai.org, Macao, China, pp. 3762–3770.
- Wang, X., Miikkulainen, R., 2020. MDEA: Malware detection with evolutionary adversarial learning. ArXiv preprint arXiv:2002.03331.
- Witten, I.H., Frank, E., 2002. Data mining: practical machine learning tools and techniques with java implementations. *ACM Sigmod Record* 31 (1), 76–77.
- Wu, C., Shi, J., Yang, Y., Li, W., 2018. Enhancing machine learning based malware detection model by reinforcement learning. In: International Conference on Communication and Network Security. ACM, Qingdao, China, pp. 74–78.
- Yan, J., Yan, G., Jin, D., 2019. Classifying malware represented as control flow graphs using deep graph convolutional neural network. In: IEEE/IFIP International Conference on Dependable Systems and Networks. IEEE, Portland, OR, USA, pp. 52–63.
- Ye, Y., Chen, L., Wang, D., Li, T., Jiang, Q., Zhao, M., 2009. Sbmds: an interpretable string based malware detection system using svm ensemble with bagging. *Journal in Computer Virology* 5 (4), 283–293.
- Ye, Y., Li, T., Adjeroh, D., Iyengar, S.S., 2017. A survey on malware detection using data mining techniques. *ACM Comput Surv* 50 (3), 1–40.
- Ye, Y., Li, T., Chen, Y., Jiang, Q., 2010. Automatic malware categorization using cluster ensemble. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, Washington, DC, USA, pp. 95–104.
- Ytisf, 2021. thezoo a live malware repo. <https://github.com/ytisf/thezoo>. Online (last accessed August 25, 2021).
- Yuan, J., Zhou, S., Lin, L., Wang, F., Cui, J., 2020. Black-box adversarial attacks against deep learning based malware binaries detection with GAN. In: European Conference on Artificial Intelligence. IOS Press, Santiago de Compostela, Spain, pp. 2536–2542.
- Zeidanloo, H.R., Tabatabaei, F., Amoli, P.V., Tajpour, A., 2010. All about malwares (malicious codes). In: Security and Management, pp. 342–348.
- Zeng, G., Qi, F., Zhou, Q., Zhang, T., Ma, Z., Hou, B., Zang, Y., Liu, Z., Sun, M., 2020. OpenAttack: an open-source textual adversarial attack toolkit. ArXiv preprint arXiv:2009.09191.
- Zhang, J., Li, C., 2019. Adversarial examples: opportunities and challenges. *IEEE Trans Neural Netw Learn Syst* 31 (7), 2578–2593.

Zhang, J., Qin, Z., Yin, H., Ou, L., Hu, Y., 2016. IRMD: malware variant detection using opcode image recognition. In: International Conference on Parallel and Distributed Systems. IEEE, Wuhan, China, pp. 1175–1180.

Zhang, L., Liu, P., Choi, Y.-H., 2020a. Semantic-preserving reinforcement learning attack against graph neural networks for malware detection. ArXiv preprint arXiv:2009.05602.

Zhang, Z., Qi, P., Wang, W., 2020. Dynamic malware analysis with feature engineering and feature learning. In: AAAI Conference on Artificial Intelligence. AAAI Press, New York, NY, USA, pp. 1210–1217.

Zhao, S., Ma, X., Zou, W., Bai, B., 2019. Deepcg: classifying metamorphic malware through deep learning of call graphs. In: International Conference on Security and Privacy in Communication Systems. Springer, Orlando, FL, USA, pp. 171–190.

Zhong, F., Cheng, X., Yu, D., Gong, B., Song, S., Yu, J., 2020. MalFox: Camouflaged adversarial malware example generation based on C-GANs against black-box detectors. ArXiv preprint arXiv:2011.01509.

Zhu, S., Zhang, Z., Yang, L., Song, L., Wang, G., 2020. Benchmarking label dynamics of virustotal engines. In: ACM SIGSAC Conference on Computer and Communications Security. ACM, Virtual Event, pp. 2081–2083.



Xiang Ling is currently a Research Associate at the Institute of Software, Chinese Academy of Sciences. He received his Ph.D. degree from Zhejiang University in 2021. His research interests include AI security, data-driven security, network/web security, and program analysis. His work has been published at top-ranked conferences and journals, like IEEE S&P, INFOCOM, TNNLS, TKDD, ICDCS, ECCV, and TOPS.



Lingfei Wu is an Engineering Manager in the Content and Knowledge Graph Group at Pinterest, where he is leading to build the next-generation Knowledge Graph to empower Pinterest recommendation/research systems across all major surfaces including Homefeed, Search, Ads, etc. He earned his Ph.D. degree in computer science from the College of William and Mary in 2016. He has published more than 90 top-ranked conference and journal papers and is a co-inventor of more than 40 filed US patents. Because of the high commercial value of his patents, he has received several invention achievement awards and has been appointed as IBM Master Inventor, class of 2020. He was the recipient of the Best Paper Award and Best Student Paper Award of several conferences such as IEEE ICC'19, AAAI workshop on DLGMA'20, and KDD workshop on DLG'19. His research has been featured in numerous media outlets, including NatureNews, YahooNews, Venturebeat, and TechTalks. He has co-organized 10+ conferences (KDD, AAAI, IEEE BigData) and is the founding co-chair for Workshops of Deep Learning on Graphs (with AAAI'21, AAAI'20, KDD'20, KDD'19, and IEEE BigData'19). He has currently served as Associate Editor for IEEE Transactions on Neural Networks and Learning Systems, ACM Transactions on Knowledge Discovery from Data and International Journal of Intelligent Systems, and regularly served as an SPC/PC member of the following major AI/ML/NLP conferences including KDD, IJCAI, AAAI, NIPS, ICML, ICLR, and ACL.



Jiangyu Zhang is a graduate student at Zhejiang University. His research focuses on binary security and data-driven security.



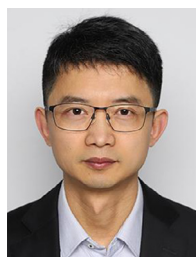
Zhenqing Qu is a graduate student at Zhejiang University. He is a member of the AAA CTF Team. His research focuses on web security and data-driven security. He has spoken at Black Hat 2022. Additionally, he has reported several severe defects to mainstream WAF vendors, which were confirmed and fixed quickly.



Wei Deng received his master degree in Computer Science and Technology from Zhejiang University at Hangzhou, China in 2022. He is now working in Tencent Security Platform Department. His research mainly focuses on network traffic security.



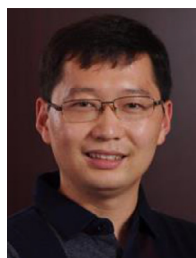
Xiang Chen received the B.Eng. and the M.Eng. degrees from Fuzhou University in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University, China. He has published papers in IEEE INFOCOM, IEEE ICNP and so on. He received the Best Paper Award from IEEE/ACM IWQoS 2021, and the Best Paper Candidate from IEEE INFOCOM 2021. His research interests include programmable networks and network security.



Yaguan Qian is a professor at the School of Big-data Science, Zhejiang University of Science and Technology. He received Ph.D. in Computer Science from Zhejiang University in 2014. He is currently a director of Hikvision-ZUST Joint Laboratory of Edge Intelligent Security. He leads a research team working on AI security, machine learning, pattern recognition, and machine vision.



Chunming Wu received a Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 1995. He is currently a Professor at the College of Computer Science and Technology, Zhejiang University. His research interests include software-defined networks, proactive network defense, network virtualization, and intelligent networks.



Shouling Ji is a ZJU 100-Young Professor in the College of Computer Science and Technology at Zhejiang University and a Research Faculty in the School of Electrical and Computer Engineering at Georgia Institute of Technology (Georgia Tech). He received a Ph.D. degree in Electrical and Computer Engineering from Georgia Institute of Technology, a Ph.D. degree in Computer Science from Georgia State University, and a B.S. (with Honors) and M.S. degrees both in Computer Science from Heilongjiang University. His current research interests include Data-driven Security and Privacy, AI Security, and Big Data Analytics. He is a member of ACM, IEEE, and CCF and was the Membership Chair of the IEEE Student Branch at Georgia State University (2012–2013). He was a Research Intern at the IBM T. J. Watson Research Center. Shouling is the recipient of the 2012 Chinese Government Award for Outstanding Self-Financed Students Abroad.



Tianyue Luo graduated from Beijing Information Science and Technology University in 2012, he is a senior engineer with the Institute of Software, Chinese Academy of Sciences, Beijing. His primary research interests include system security, vulnerability detection, and open source software security.



Jingzheng Wu received the Ph.D. degree in computer software and theory from the Institute of Software, Chinese Academy of Sciences, Beijing, in 2012. He is a research professor with the Institute of Software, Chinese Academy of Sciences, Beijing. His primary research interests include system security, vulnerability detection, and covert channels.



Yanjun Wu received the Ph.D. degree in computer software and theory from the Institute of Software, Chinese Academy of Sciences, Beijing, in 2006. He is currently a Research Professor with the Institute of Software, Chinese Academy of Sciences, Beijing. His current research interests include operating systems and system security.