MalNet: A Large-Scale Image Database of Malicious Software

Scott Freitas*
safreital@gmail.com
Georgia Institute of Technology
Atlanta, GA, USA

Rahul Duggal
rahamduggal@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Duen Horng Chau
polo@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

Computer vision is playing an increasingly important role in automated malware detection with the rise of the image-based binary representation. These binary images are fast to generate, require no feature engineering, and are resilient to popular obfuscation methods. Significant research has been conducted in this area, however, it has been restricted to small-scale or private datasets that only a few industry labs and research teams have access to. This lack of availability hinders examination of existing work, development of new research, and dissemination of ideas. We release MalNet-IMAGE, the largest public cybersecurity image database, offering 24x more images and 70x more classes than existing databases (available at https://malnet.org). MalNet-IMAGE contains over 1.2 million malware images—across 47 types and 696 families—democratizing image-based malware capabilities by enabling researchers and practitioners to evaluate techniques that were previously reported in propriety settings. We report the first million-scale malware detection results on binary images. MalNet-IMAGE unlocks new and unique opportunities to advance the frontiers of machine learning, enabling new research directions into vision-based cyber defenses, multi-class imbalanced classification, and interpretable security.

CCS CONCEPTS

• Security and privacy → Malware and its mitigation; • Computing methodologies → Neural networks; Supervised learning by classification.

KEYWORDS

Dataset, deep learning, supervised learning, computer vision, imbalanced classification, cybersecurity, malware detection, binary image

1 INTRODUCTION

Attack campaigns from criminal organizations and nation state actors are one of the most powerful forms of disruption, costing the U.S. economy as much as $109 billion a year [40]. These cyber attacks are highly sophisticated, targeting governments and large-scale enterprises to interrupt critical services and steal intellectual property [18]. Defending against these attacks requires the development of strong antivirus tools to identify new variants of malicious software before they can infect a network. Unfortunately, as a majority of newly identified malware is polymorphic in nature, where a few subtle source code changes result in significantly different compiled code (e.g., instruction reordering, branch inversion, register allocation) [15, 53], the predominant signature-based form of malware detection is rendered inert [46].

To combat these issues, the cybersecurity industry [7] has turned to image-based malware representations as they are quick to generate, require no feature engineering, and are resilient to common obfuscation techniques (e.g., section encryption [37], file packing [38]). For all of these reasons, image-based malware detection and classification research has surged in popularity. Unfortunately, a majority of this research uses small-scale or private data repositories, making it increasingly difficult to characterize and differentiate existing work, develop new research methodologies, and disseminate new ideas [7, 9, 16, 19–21, 32, 33, 37, 38, 42]. To address these issues, we constructed MalNet-IMAGE, the first large-scale ontology of malicious software images.

1.1 Contributions

1. Largest Cybersecurity Image Database. MalNet-IMAGE contains over 1.2 million software images across a hierarchy of 47 types and 696 families, enabling researchers and practitioners to conduct experiments on an industry scale dataset, and evaluate techniques...
that were previously reported in propietary settings. Compared to the next large public database [39], MALNET-IMAGE offers 24× more images and nearly 70× more classes (see Table 1). We report the first public large-scale malware detection and classification results on binary images, where we are able to detect malicious files with an AUC of 0.94 and classify them across 47 types and 696 families with a macro-F1 score of 0.49 and 0.45, respectively.

2. Permissive Licensing & Open Source Code. We release MALNET-IMAGE with a CC-BY license, allowing researchers and practitioners to share and adapt the database to their needs. We open-source the code to create the images and run the experiments on Github.

3. Visual Exploration Without Downloading. We develop MALNET-IMAGE EXPLORER, an image exploration and visualization tool that enables researchers and practitioners to easily study the data without installation or download. MALNET-IMAGE EXPLORER is available online at: https://mal-net.org.

4. Community Impact. MALNET-IMAGE offers new and unique opportunities to advance the frontiers of cybersecurity research. In particular, MALNET-IMAGE offers researchers a chance to study imbalanced classification on a large-scale cybersecurity database with a natural imbalance ratio of 16, 901× (see Figure 1); and explore explainability research in a high impact domain, where it is critical that security analysts can interpret and trust the model.

2 ADVANCING THE STATE-OF-THE-ART

Aside from MALNET-IMAGE, there are only two publicly available binary-image based cybersecurity datasets— Malimg [38] and Virus-MNIST [39]—containing 9,458 images across 25 classes, and 51,880 images across 10 classes, respectively. In surveying the malware detection and classification literature [3–8, 16, 19–21, 26, 28, 29, 33–35, 38, 43, 48, 54], we observed that almost all experiments were conducted on small-scale or private data. As the field advances, large-scale public databases are necessary to develop the next generation of algorithms. In Table 1, we compare MALNET-IMAGE with other public and private cybersecurity image datasets. We find that that MALNET-IMAGE offers 24× more images and 70× the classes, compared to the largest alternative public binary image database (Virus-MNIST [39]); and 479,800 more images and 694 more classes than the largest private database (Stamina [7]). We do not compare against repositories of malicious binaries such as AndroZoo [30], AMD [50], Microsoft-BIG [45], Malicia [36], VirusShare, and VirusTotal in this discussion, as none of them have images available to use.

Security Implications. With the release of MALNET-IMAGE, researchers will now have access to a critical resource to develop advanced, image-based malware detection and classification algorithms. Like most open data resources, there is a potential for misuse by malicious actors who aim to craft new variants to evade detection. We believe MALNET-IMAGE’s contribution to the research community significantly outweighs such risk.

2.1 Constructing MalNet-Image
MALNET-IMAGE is an ambitious project to collect and process over 1.2 million binary images, and is a major extension to the graph representation learning database MALNET [17], offering significant new malware detection capabilities. Below, we describe the provenance and construction of MALNET-IMAGE.

Collecting Candidate Images. We construct MALNET-IMAGE using the Android ecosystem due to its large market share [41], easy accessibility [30] and diversity of malicious software [1]. With the generous permission of AndroZoo [2, 30], we collected 1,262,024 Android APK files, specifically selecting APKs containing both a family and type label obtained from Euphony [25], a state-of-the-art malware labeling system that aggregates and learns from the labeling results of up to 70 antivirus vendors from VirusTotal [49].

Processing the Images. The first step in constructing the image representation was to extract the DEX file (bytecode) from each Android APK. The extracted DEX file was then converted into a 1D array of 8 bit unsigned integers. Each entry in the array is in the range [0, 255] where 0 corresponds to a black pixel and 255 a white pixel. We then convert each 1D byte array into a 2D array using standard linear plotting where the width of the image is fixed

Figure 2: Left: Android DEX file structure, composed of three major components—(1) header, (2) ids, and (3) data. Right: binary image representation of the DEX file.
and the height is allowed to vary based on the file size. We use the same image width proposed on the seminal work [37] (and follow up work [11, 27, 44, 51]), and scale each image to 256 × 256 using a standard Lanczos filter from the Pillow library. Finally, we color each byte according to its use, adding a layer of semantic information on top of the raw bytecode. While a variety of techniques can be used to encode semantic information into the image, there is currently no accepted standard. We follow [20] and assign each byte to a particular RGB channel depending on its position in the DEX file structure—(i) header, (ii) identifiers and class definitions, and (iii) data (see Figure 2). Distributed across Google Cloud’s General-purpose (N2) machine with 16 cores running 24 hours a day, this process took approximately a week. We release the source code used to process the APKs on Github.

**MalNet-Image Tiny.** We construct **MALNET-IMAGE TINY**, containing 61,201 training, 8,743 validation and 17,486 test images, for type level classification experiments using the 4 largest types in **MALNET-IMAGE**. The goal of **MALNET-IMAGE-TINY** is to enable users to rapidly prototype new ideas, since it requires only a fraction of the time needed to train a new model. **MALNET-IMAGE TINY** is released alongside the full dataset at https://mal-net.org.

### 3 MALNET-IMAGE APPLICATIONS

**MALNET-IMAGE** offers new and unique opportunities to advance the frontiers of cybersecurity research. As examples, we show three exciting new applications made possible by the **MALNET-IMAGE** database—(1) as a state-of-the-art cybersecurity image benchmark in Section 3.1; (2) as the first large-scale public analysis of malicious software detection using binary images in Section 3.2; and (3) how to categorize high-risk malware threats (e.g., is this Ransomware or Spyware?) in Section 3.3. Then, in Section 3.4 we highlight new research directions enabled by **MALNET-IMAGE**.

**Application Setup.** We divide **MALNET-IMAGE** into three stratified sets of data, with a training-validation-test split of 70-10-20 respectively; repeated for both type and family labels (suggested splits available at https://mal-net.org). In addition, we conduct malware detection experiments by grouping all 46 malicious software images into one type while the benign type maintains its original label. We evaluate 3 common architectures—ResNet [22], DenseNet [24] and MobileNet [23], based on its macro-F1 score, as is typical for highly imbalanced datasets [12–14, 17]. Each model is trained for 100 epochs using cross entropy loss (unless specified otherwise) and an Adam optimizer on an Nvidia DGX-1 containing 8 V100 GPUs and 512GB of RAM using Keras with a Tensorflow backend.

#### 3.1 Application 1: Benchmarking Techniques

Leveraging the unprecedented scale and diversity of **MALNET-IMAGE**, we evaluate numerous malware detection and classification techniques that have previously been studied using only private or small-scale databases. Specifically, we evaluate recent techniques including: (a) semantic information encoding via colored channels, (b) model architecture, (c) imbalanced classification techniques, and (d) the performance of **MALNET-IMAGE TINY**, a small-scale version of **MALNET-IMAGE**. We detail the setup, results, and analysis of each experiment below.
types benefit the most from class reweighting, where the ‘Click’ type (113 examples), sees its F1 score rise from 0 to 0.91. On the other hand, focal loss shows no improvement over the baseline model, likely due to its design for use in dense object detectors like R-CNN. Going forward, all experiments use cross entropy loss with class reweighting due to the strong improvement in smaller malware types.

**MALNET-IMAGE Tiny Performance.** We train a ResNet18 model on grayscale images using cross entropy loss and class reweighting, and achieve a macro-F1 score of 0.65. Compared to full dataset, the macro-F1 score is significantly higher (0.65 vs 0.49); which is unsurprising since the largest 4 types contained a significant proportion of the image diversity (based on the number of families), resulting in an easier classification task.

**Limitations.** Methods that work well on other datasets may not work well on MALNET-IMAGE due to structural differences in the images; vice-versa, methods that work on MALNET-IMAGE may not transfer well to other datasets. We hope this work inspires new research in the binary image domain, enabling the development of methods that generalize across key domains such as cybersecurity.

### 3.2 Application 2: Malware Detection

Researchers and practitioners can now conduct malware detection experiments on an industry scale dataset, evaluating things that were previously reported in propriety settings. Using the model selected in Section 3.1—a ResNet18 model trained on grayscale images using cross entropy loss and class reweighting—we perform an in-depth analysis of this highly imbalanced detection problem containing 1,182 malicious and 79 benign images. We find that the model is able to obtain a strong macro-F1 score of 0.86, macro-precision of 0.89 and a macro-recall of 0.84. We further study the model’s detection capabilities by analyzing its ROC curve, where the model achieves an AUC score of 0.94, and is able to identify 84% of all malicious files with a false positive rate of 10% (a common threshold used in security [7]). This first of its kind analysis allows researchers insight into malware detection that is usually restricted to handful of industry labs.

MALNET-IMAGE also opens new opportunities in the nascent and promising research direction of analyzing attention maps to interpret malware detection results. Yakura et al. [51, 52] showed that specific byte sequences found in the attention map closely correlate with malicious code payloads. We evaluate the potential of attention maps on MALNET-IMAGE using the popular Grad-Cam [47] technique to highlight regions of interest across 3 types of malware and benignware in Figure 3. In the malware images (left three), we see the attention map is focused on thin regions of bytecode in the data section (where malicious payloads are often stored), while in the benign images (right side) the attention map is dispersed across the larger data region. This type of visual analysis can significantly reduce the amount of time and effort required to manually investigate a file by guiding security analysts to suspicious regions of the bytecode.

### 3.3 Application 3: Malware Classification

MALNET-IMAGE opens up new research into binary images as a tool for multi-class malware classification (e.g., is this file Ransomware or Spyware?). Using the model selected in Section 3.1, we perform an in-depth analysis of its multi-class classification capability across 47 types and 696 families of malware. We find the model is able to classify the malware type and malware family with a macro-F1 score of 0.49 and 0.45, respectively.

In Figure 4, we conduct an in-depth analysis into type level classification performance through a confusion matrix heatmap. A dark diagonal indicates strong classifier performance, where a dark off-diagonal entry indicates poor performance. Each square in the diagonal indicates the percent of examples correctly classified for a particular malware type, and each off-diagonal row entry indicates the percent of incorrectly classified examples for a particular malware type. We find that four types of malware comprise the majority of misclassifications: Adware, Benign, Riskware, and Trojan. Unsurprisingly, these are the 4 largest types of malware (based on the number of images in each class), indicating the strong effect that data imbalance has in the malware classification process. In addition, the heatmap can be used to identify potential naming disagreements between vendor labels (e.g., “adware” and “adsware”), serving as a basis for merging certain types of malware. To the best of our knowledge, this is the first public large-scale analysis of malware classification, providing a new state-of-the-art benchmark to compare against.

### 3.4 Enabling New Research Directions

The scale and diversity of MALNET-IMAGE opens up new exciting research opportunities to the ML and security communities. Below, we present 3 promising directions (R1–R3).

**R1. Advancing Vision Based Cybersecurity Research.** Research into developing image-based malware detection and classification algorithms has recently surged across industry (e.g., Intel-Microsoft collaboration on Stamina [7], security companies [20, 39]) and academia [3–6, 8, 16, 19, 21, 26, 28, 29, 33–35, 43, 48, 54]. However, existing public datasets contain only a handful of classes and thousands of images, and as the field advances, larger and more challenging datasets are needed for
Figure 4: Malware classification results using confusion matrix heatmap (classes in descending order of number of samples). We analyze type level classification performance, where a dark diagonal indicates strong performance, and a dark off-diagonal indicates poor performance. Each square in the diagonal indicates the percent of examples correctly classified for a particular malware type, and each off-diagonal entry indicates the percent of incorrectly classified examples for a particular type.
the next generation of models. With the release of MALNET-IMAGE, researchers have access to a critical resource to develop and benchmark advanced image-based malware detection and classification algorithms, previously restricted to a few industry labs and research teams.

R2. Extending Imbalanced Classification to a New Domain. Only preliminary work has studied binary-image malware classification under data imbalance [54] due to the limited number of classes and images available in public datasets. As a result, it is unknown whether many techniques may generalize to the binary-image domain, and how they will perform in highly imbalanced classification scenarios. We take a first step in this direction in Section 3.1, where we show that classes containing only a few examples typically underperform relative to their more populous counterparts—highlighting the significant challenge of imbalanced classification in the cybersecurity domain. By releasing MALNET-IMAGE, one of the largest naturally imbalanced databases to date, we hope to foster new interest in this important research area, enabling the machine learning community to impact and generalize across domains.

R3. Interpretable Cybersecurity Research. Preliminary research has demonstrated the importance of attention mechanisms in binary-image malware classification, where extracted regions can provide strong indicators to human analysts, helping guide them to suspicious parts of the bytecode for additional analysis [51, 52]. This includes recent research in salience based methods that automatically discover concepts, helping to identify correlated regions of bytecode [55]. Prior to MALNET-IMAGE, researchers were limited to a small number of malicious families and types, hindering their ability to conduct large-scale explainability studies. With MALNET-IMAGE’s nearly 700 classes, researchers can explore a wide variety of malicious software, enabling new breakthroughs and discoveries. For example, researchers might discover that new types of visualization and sense-making techniques are needed to accurately summarize large volumes of binary-image data to enhance security analysts decision making capabilities.

4 CONCLUSION

Computer vision research into binary-image malware detection and classification is a crucial tool in protecting enterprise networks and governments from cyber attacks seeking to interrupt critical services and steal intellectual property. Leveraging MALNET-IMAGE’s scale and diversity—containing 1, 262, 024 binary images across a hierarchy of 47 types and 696 families—researchers and practitioners can now conduct experiments that were previously restricted to a few industry labs and research teams. We hope MALNET-IMAGE becomes a central resource for a broad range of research into vision-based cyber defenses, multi-class imbalanced classification, and interpretable security.

ACKNOWLEDGMENTS

We thank Kevin Allix and AndroZoo colleagues for generously allowing us to use their data in this research; this work was in part supported by NSF grant IIS-1563816, CNS-1704701, GRFP (DGE-1650044), a Raytheon research fellowship, and an IBM fellowship.

REFERENCES

[1] 2018. Nokia Threat Intelligence Report. Network Security (2018). https://doi.org/10.1016/S1353-4858(18)30122-3
[2] Kevin Allix, Tegawendé F Bissyandé, Jacques Klein, and Yves Le Traon. 2016. Androzoo: Collecting millions of android apps for the research community. In 2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR). IEEE, 468–471.
[3] Ahmad Azab and Mahmoud Khasawneh. 2020. MSC: Malware Spectrogram Image Classification. IEEE Access 8 (2020), 102007–102021.
[4] Roland Burks, Kari Aminal Islam, Yan Lu, and Jiang Li. 2019. Data Augmentation with Generative Models for Improved Malware Detection: A Comparative Study. In UEMCON. IEEE, 0660–0665.
[5] Ferhat Ozgur Catak, Javed Ahmed, Kevser Sahinbas, and Zahid Hussain Khan. 2020. Data Augmentation Based Malware Detection using Convolutional Neural Networks. arXiv preprint arXiv:2010.01862 (2020).
[6] Li Chen. 2018. Deep transfer learning for static malware classification. arXiv preprint arXiv:1812.07606 (2018).
[7] Li Chen, Ravi Sahita, Jugal Parikh, and Marc Marino. [n.d.]. Stamina: scalable deep learning approach for malware classification. Intel White Paper (n.d.).
[8] Sunoh Choi, Sungwook Jang, Youngsoo Kim, and Jonghyun Kim. 2017. Malware detection using malware image and deep learning. In ICFP. IEEE, 1193–1195.
[9] Gregory Conti, Sergio Bratus, Anna Subhina, Andrew Lichtenberg, Roy Ragdale, Robert Perez-Alemany, Benjamin Sangster, and Matthew Sagan. 2010. A visual study of primitive binary fragment types. Black Hat USA (2010).
[10] Yun Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In CVPR. 9268–9277.
[11] Zhizhuo Cui, Fei Xue, Xinjian Cai, Yang Cao, Gai-ge Wang, and Jinjun Chen. 2018. Detection of malicious code variants based on deep learning. IEEE Transactions on Industrial Informatics 14, 7 (2018), 3187–3196.
[12] Rahul Duggal, Scott Freitas, Sunny Dhamnani, Duen Horng Chau, and Jimeng Sun. 2021. HAR: Hardness Aware Reweighting for Imbalanced Datasets. In 2021 IEEE International Conference on Big Data (Big Data). IEEE, 735–745.
[13] Rahul Duggal, Scott Freitas, Sunny Dhamnani, Duen Horng, Jimeng Sun, et al. 2020. ELF: An Early-Exiting Framework for Long-Tailed Classification. arXiv:2006.11979 (2020).
[14] Rahul Duggal, Scott Freitas, Cao Xiao, Duen Horng Chau, and Jimeng Sun. 2020. REST: Robust and Efficient Neural Networks for Sleep Monitoring in the Wild. In Proceedings of The Web Conference 2020. 1704–1714.
[15] Thomas Dullien and Rolf Rolles. 2005. Graph-based comparison of executable objects. SSTC (2005).
[16] Yong Fang, Yangchen Gao, Fan Jing, and Lei Zhang. 2020. Android Malware Familial Classification Based on DEX File Section Features. IEEE Access 8 (2020), 10614–10627.
[17] Scott Freitas, Yuxiao Dong, Joshua Neil, and Duen Horng Chau. 2020. A Large-Scale Database for Graph Representation Learning. arXiv preprint arXiv:2011.07682 (2020).
[18] Scott Freitas, Andrew Wicker, Duen Horng Chau, and Joshua Neil. 2020. D2M: Dynamic Defense and Modeling of Adversarial Movement in Networks. In Proceedings of the 2020 SIAM International Conference on Data Mining. SIAM.
[19] Jianwen Fu, Jinfeng Xue, Yong Wang, Zhanyan Liu, and Chun Shan. 2018. Malware visualization for fine-grained classification. IEEE Access 6 (2018), 14510–14523.
[20] Joedy Gennissen, Lorenzo Cavallaro, Veelasha Moonamsy, and Lejla Batina. 2017. Gamut: Sifting through Images to Detect Android Malware. Kyoung Soo Han, Jae Hyun Lim, Bojoong Kang, and Eul Gyu Im. 2015. Malware analysis using visualized images and entropy graphs. International Journal of Information Security 14, 1 (2015), 1–14.
[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.
[22] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017).
[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2016. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.
