Supplementary Materials

Leveraging intelligent multimodal fusion for few-shot malware classification

Ying Ren[1], Ziyu Liu[2], Junbo Wang[3], Peng Wang[4]

[1]Department of Outpatient, West China Hospital, Sichuan University, Chengdu 610065, Sichuan, China.
[2]The Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou 310058, Zhejiang, China.
[3]College of Software Engineering, Sichuan University, Chengdu 610065, Sichuan, China.
[4]College of Computer Science, Sichuan University, Chengdu 610065, Sichuan, China.

Correspondence to: Prof. Ying Ren, Department of Outpatient, West China Hospital, Sichuan University, Chengdu 610065, Sichuan, China. E-mail: zq156157@163.com

DATASETS

VirusShare-M

This dataset comprises Windows PE malware samples downloaded from VirusShare website [1]. We adopted the labeling methodology described in [2,3], utilizing VirusTotal [4] to scan these malware samples and obtain reports in JSON format. These reports contained categorizations of the malware samples from diferent antivirus engines. Due to discrepancies in labeling across various antivirus engines, we employed the AVClass tool [5] to standardize the labels. Ultimately, we identified 127 malware families, each containing more than 20 samples. From each family, we randomly selected 20 samples to form the final dataset. We extracted grayscale image features and API call sequence features for each sample. The resulting dataset is divided into *Dtrain*, *Dval*, and D^{test} in a ratio of 87:20:20, with 20 samples per class. Detailed class names are provided in Supplementary Table 1.

Supplementary Table 1. Details of class names for VirusShare-M Dataset

Subset	Owned class names	Count
Dtrain	acda, adclicer, airinstaller, antavmu, autoit, badur, bundlore, c99shell, cidox, conficker, cpllnk, darbyen, darkkomet, dealply, decdec, delbar, dlhelper, domaiq, downloadadmin, downloadsponsor, egroupdial, extenbro, fbjack, fearso, fosniw, fsysna, gamevance, gepys, getnow, goredir, hicrazyk, hidelink, hijacker, hiloti, ibryte, iframeinject, includer, installmonetizer, ipamor, ircbot, jeefo, jyfi, kovter, linkury, lipler, llac, loadmoney, lunam, microfake, midia, mikey, mydoom, nimda, nitol, outbrowse, patchload, pullupdate, qhost, qqpass, reconyc, redir, refresh, scarsi, scrinject, shipup, simbot, softcnapp, softonic, softpulse, somoto, staser, toggle, trymedia, unruy, urelas, vilsel, vittalia, wabot, webprefix, wonka, xorer, xtrat, yoddos, zapchast, zegost, zvuzona, zzinfor	87
Dval	1clickdownload, bettersurf, black, buterat, directdownloader, faceliker, fujacks, icloader, iframeref, kido, kykymber, lineage, linkular, loring, msposer, pirminay, pykspa, soft32downloader, startp, zeroaccess	20
Dtest	4shared, banload, blacole, browsefox, downloadassistant, fakeie, firseria, gator, inor, installerex, instally, psyme, refroso, sefnit, sytro, urausy, vtflooder, wajam, windef, zbot	20

LargePE-M

This dataset consists of samples collected from a large repository of Windows PE malware maintained by our laboratory, encompassing over 222,700 samples from more than 500 malware families. The majority of samples were downloaded from VirusTotal [4], and their categorization followed the methodology described in [2]. It is noteworthy that during the construction of API call sequences using the method described in Section 3.2, some malware samples failed to execute within the sandbox environment. These samples were excluded from the dataset. We selected categories with more than 10 samples from which we could successfully extract API sequence features. The dataset was then split into *Dtrain*, *Dval*, and D^{test} in a ratio of 34:11:13, with each class containing between 10 and 20 samples (randomly selecting 20 samples for classes with more than 20 samples). In total, 981 samples were included, with 570 for training, 177 for validation, and 234 for testing. Detailed class names are provided in Supplementary Table 2.

Supplementary Table 2. Details of class names for LargePE-M Dataset

Subset	Owned class names	Count
Dtrain	backdoor.win32.ceckno, backdoor.win32.gobot, trojan.win32.sadenav, backdoor.win32.singu, trojan-mailfinder.win32.blen, trojan.win32.midgare, trojan.win32.humor, trojan-downloader.win32.vb, trojan-dropper.win32.small, trojan-spy.win32.sckeylog, trojan-spy.win32.flux, trojan-dropper.win32.vb, trojan-spy.win32.banker, backdoor.win32.prorat, dos.win32.vb, trojan-banker.win32.banpaes, hacktool.win32.vb, trojan.win32.obfuscated, trojan-spy.win32.flystudio, trojan-spy.win32.winspy, trojan-clicker.win32.delf, backdoor.win32.bifrose, backdoor.win32.dsbot, trojan-psw.win32.immultipass, net-worm.win32.kolab, virus.win32.hllw, virus.win32.xorer, worm.win32.vb, trojan-gamethief.win32.tibia, trojan-dropper.win32.js, trojan-spy.win32.vb, trojan.win32.bholamp, trojan.win32.diamin, trojan.win32.antiav	34
Dval	trojan-psw.win32.tibia, trojan-downloader.win32.swizzor, trojan-spy.win32.zbot, trojan- clicker.win32.osewlone, email-worm.win32.joleee, trojan-spy.win32.keylogger, trojan.win32.favadd, backdoor.win32.frauder, backdoor.win32.hupigon, trojan-spy.win32.bzub, trojan-clicker.win32.small	11
Dtest	trojan-spy.win32.banbra, trojan-clicker.win32.vb, backdoor.win32.darkmoon, trojan- banker.win32.banker, trojan.win32.regrun, trojan-psw.win32.maran, backdoor.win32.girlinred, trojan-spy.win32.ayolog, virus.win32.vb, worm.win32.downloader, backdoor.win32.rshot, trojan- psw.win32.vb, trojan-dropper.win32.parsi	13

Dataset Limitations: The VirusShare-M and LargePE-M datasets are constructed from publicly available malware repositories and may not fully represent the diversity and evolution of real-world malware. Additionally, the API sequences are extracted using a sandbox environment, which may fail to capture the full behavior of evasive malware that detects virtualized environments. Researchers should exercise caution when generalizing results from these datasets to operational settings and consider augmenting evaluation with more diverse and temporally stratified datasets.

REFERENCES

1. VirusShare. VirusShare; 2011. https://virusshare.com.
2. Tang Z, Wang P, Wang J. ConvProtoNet: Deep prototype induction towards better class representation for few-shot malware classification. *Applied Sciences* 2020;10:2847.
3. Wang P, Tang Z, Wang J. A novel few-shot malware classification approach for unknown family recognition with multi-prototype modeling. *Computers & Security* 2021;106:102273.
4. VirusTotal. Virustotal-free online virus, malware and url scanner; 2011. https://www.virustotal.com.
5. Sebastián M, Rivera R, Kotzias P, Caballero J. AVclass: A Tool for Massive Malware Labeling. In: Research in Attacks, Intrusions, and Defenses - 19th International Symposium, RAID 2016, Paris, France, September 19-21, 2016, Proceedings. vol. 9854 of Lecture Notes in Computer Science; 2016. pp. 230–53. Available from: https://doi.org/10.1007/978-3-319-45719-2-11.