Malware Classification based on Static Analysis

Malware Classification based on Static Analysis

- Write the code to implement the malware classification method based on the below paper:

  Nataraj, Lakshmanan, et al. "Malware images: visualization and automatic classification." *Proceedings of the 8th international symposium on visualization for cyber security*. 2011.

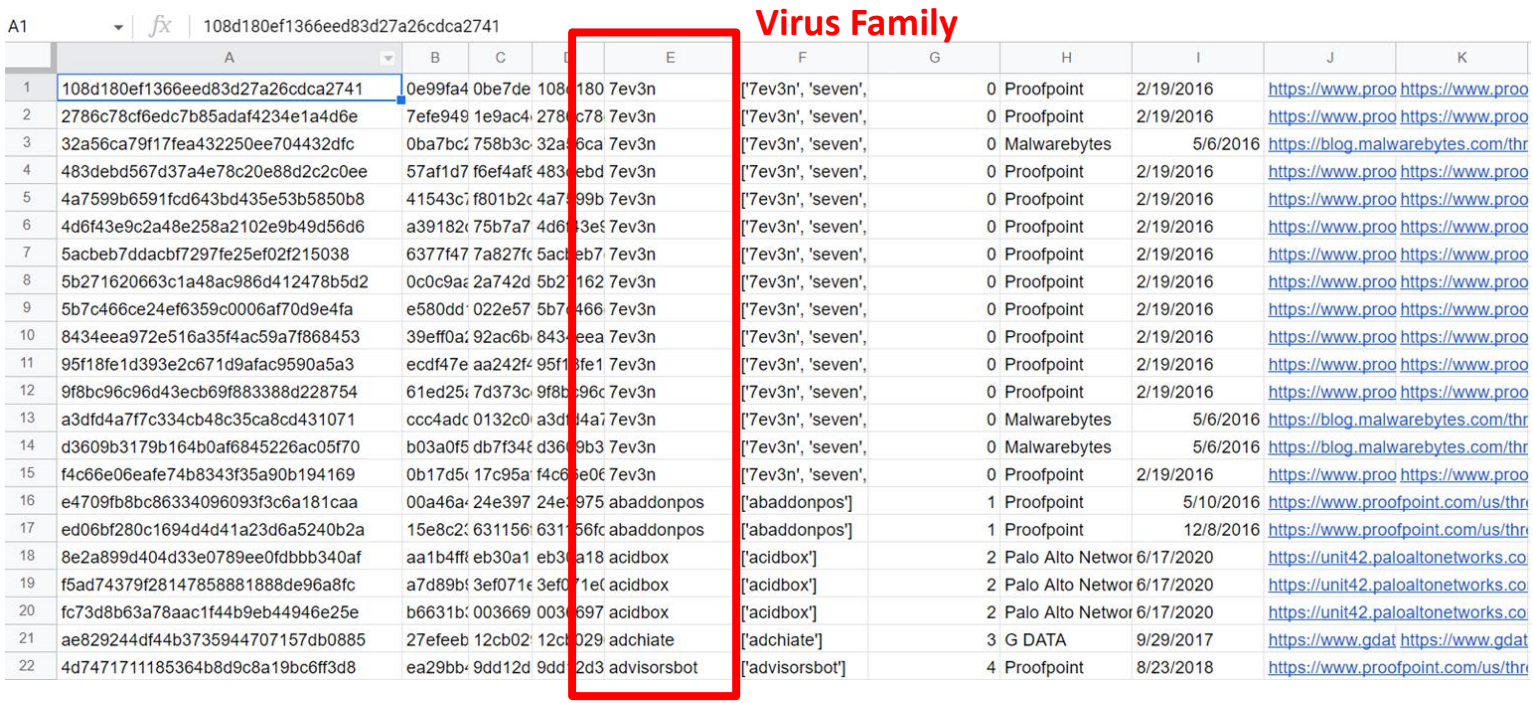
  - The feature extraction methods (GIST,…), and machine learning methods (CNN,…) are not limited. You can try different ones, and make some *comparison*.

    [Some Reference]

    - Kumar, Nitish, and Toshanlal Meenpal. "Texture-based malware family classification." *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019.
    - Bensaoud, Ahmed, Nawaf Abudawaood, and Jugal Kalita. "Classifying malware images with convolutional neural network models." *International Journal of Network Security* 22.6 (2020): 1022-1031.
    - Kiger, John, Shen-Shyang Ho, and Vahid Heydari. "Malware Binary Image Classification Using Convolutional Neural Networks." *International Conference on Cyber Warfare and Security*. Vol. 17. No. 1. Academic Conferences International Limited, 2022.
    - Vasan, Danish, et al. "Image-Based malware classification using ensemble of CNN architectures (IMCEC)." *Computers & Security* 92 (2020): 101748.

Hints

1. Malware dataset: Collect samples from VirusShare or MOTIF Dataset (with labeled virus hash value)
   - **Be Careful. You may need to do it in the Virtual Machine (VMWare, VirtualBox, …).**
   - **Do not use pre-existing transformed datasets, such as the MalImg dataset.** You need to perform the image transformation using your own code.



A1 | fx | 108d180ef1366eed83d27a26cdca2741

	A	B	C	D	E	F	G	H	I	J	K
1	108d180ef1366eed83d27a26cdca2741	0e99fa4	0be7de	108c180	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
2	2786c78cf6edc7b85adaf4234e1a4d6e	7efe949	1e9ac4	2786c78	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
3	32a56ca79f17fea432250ee704432dfc	0ba7bc2	758b3c	32a56ca	7ev3n	['7ev3n', 'seven',	0	Malwarebytes	5/6/2016	https://blog.malwarebytes.com/thr	
4	483debd567d37a4e78c20e88d2c2c0ee	57af1d7	f6ef4af8	483debd	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
5	4a7599b6591fcd643bd435e53b5850b8	41543c7	f801b2c	4a7599b	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
6	4d6f43e9c2a48e258a2102e9b49d56d6	a39182c	75b7a7	4d6f43e	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
7	5acbeb7ddacbf7297fe25ef02f215038	6377f47	7a827fc	5acbeb7	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
8	5b271620663c1a48ac986d412478b5d2	0c0c9a2	2a742d	5b27162	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
9	5b7c466ce24ef6359c0006af70d9e4fa	e580dd	022e57	5b7c466	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
10	8434eea972e516a35f4ac59a7f868453	39eff0a2	92ac6b	8434eea	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
11	95f18fe1d393e2c671d9afac9590a5a3	ecdf47e	aa242f4	95f18fe1	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
12	9f8bc96c96d43ecb69f883388d228754	61ed25a	7d373c	9f8bc96c	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
13	a3dfd4a7f7c334cb48c35ca8cd431071	ccc4adc	0132c0	a3dfd4a7	7ev3n	['7ev3n', 'seven',	0	Malwarebytes	5/6/2016	https://blog.malwarebytes.com/thr	
14	d3609b3179b164b0af6845226ac05f70	b03a0f5	db7f348	d3609b3	7ev3n	['7ev3n', 'seven',	0	Malwarebytes	5/6/2016	https://blog.malwarebytes.com/thr	
15	f4c66e06eafe74b8343f35a90b194169	0b17d5c	17c95a	f4c66e06	7ev3n	['7ev3n', 'seven',	0	Proofpoint	2/19/2016	https://www.proo	https://www.proo
16	e4709fb8bc86334096093f3c6a181caa	00a46a4	24e397	24e3975	abaddonpos	['abaddonpos']	1	Proofpoint	5/10/2016	https://www.proofpoint.com/us/thr	
17	ed06bf280c1694d4d41a23d6a5240b2a	15e8c23	631156	631156fc	abaddonpos	['abaddonpos']	1	Proofpoint	12/8/2016	https://www.proofpoint.com/us/thr	
18	8e2a899d404d33e0789ee0fdbbb340af	aa1b4ff8	eb30a1	eb30a18	acidbox	['acidbox']	2	Palo Alto Networ	6/17/2020	https://unit42.paloaltonetworks.co	
19	f5ad74379f28147858881888de96a8fc	a7d89b9	3ef071e	3ef071e0	acidbox	['acidbox']	2	Palo Alto Networ	6/17/2020	https://unit42.paloaltonetworks.co	
20	fc73d8b63a78aac1f44b9eb44946e25e	b6631b3	003669	003c697	acidbox	['acidbox']	2	Palo Alto Networ	6/17/2020	https://unit42.paloaltonetworks.co	
21	ae829244df44b37359447071 57db0885	27efeeb	12cb02	12cb029	adchiate	['adchiate']	3	G DATA	9/29/2017	https://www.gdat	https://www.gdat
22	4d7471711185364b8d9c8a19bc6ff3d8	ea29bb4	9dd12d	9dd12d3	advisorsbot	['advisorsbot']	4	Proofpoint	8/23/2018	https://www.proofpoint.com/us/thr	

Hints

2. Transform each malware to a grey image.

- A given malware binary is read as a vector of 8 bit unsigned integers and then organized into a 2D array. This can be visualized as a gray scale image in the range [0,255] (0: black, 255: white). The width of the image is fixed and the height is allowed to vary depending on the file size (Fig. 1). Tab. 1 gives some recommended image widths for different file sizes based on empirical observations.

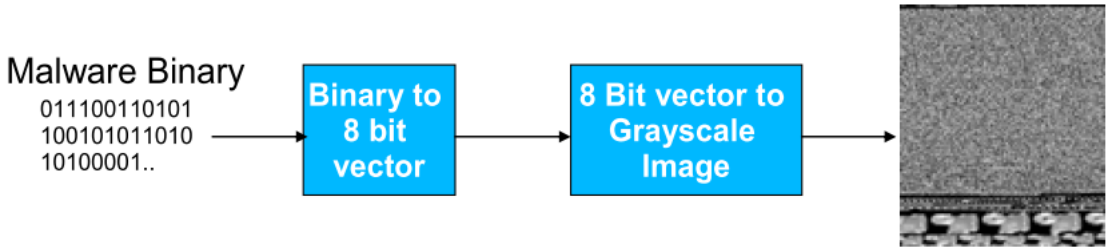


Fig.1 Visualizing Malware as an Image

Tab. 1: Image Width for Various File Sizes

File Size Range	Image Width
<10 kB	32
10 kB – 30 kB	64
30 kB – 60 kB	128
60 kB – 100 kB	256
100 kB – 200 kB	384
200 kB – 500 kB	512
500 kB – 1000 kB	768
>1000 kB	1024

- There are several feature extraction methods: GIST, HOG, LBP, … (Some methods even try to eat the image of the whole file)
- There are many machine learning methods: CNN, SVM, …
- You could make some comparison, if possible.

3. Training and testing by machine learning model for detecting malware (grey image).

■ Programming Language is not limited.

Malware Images

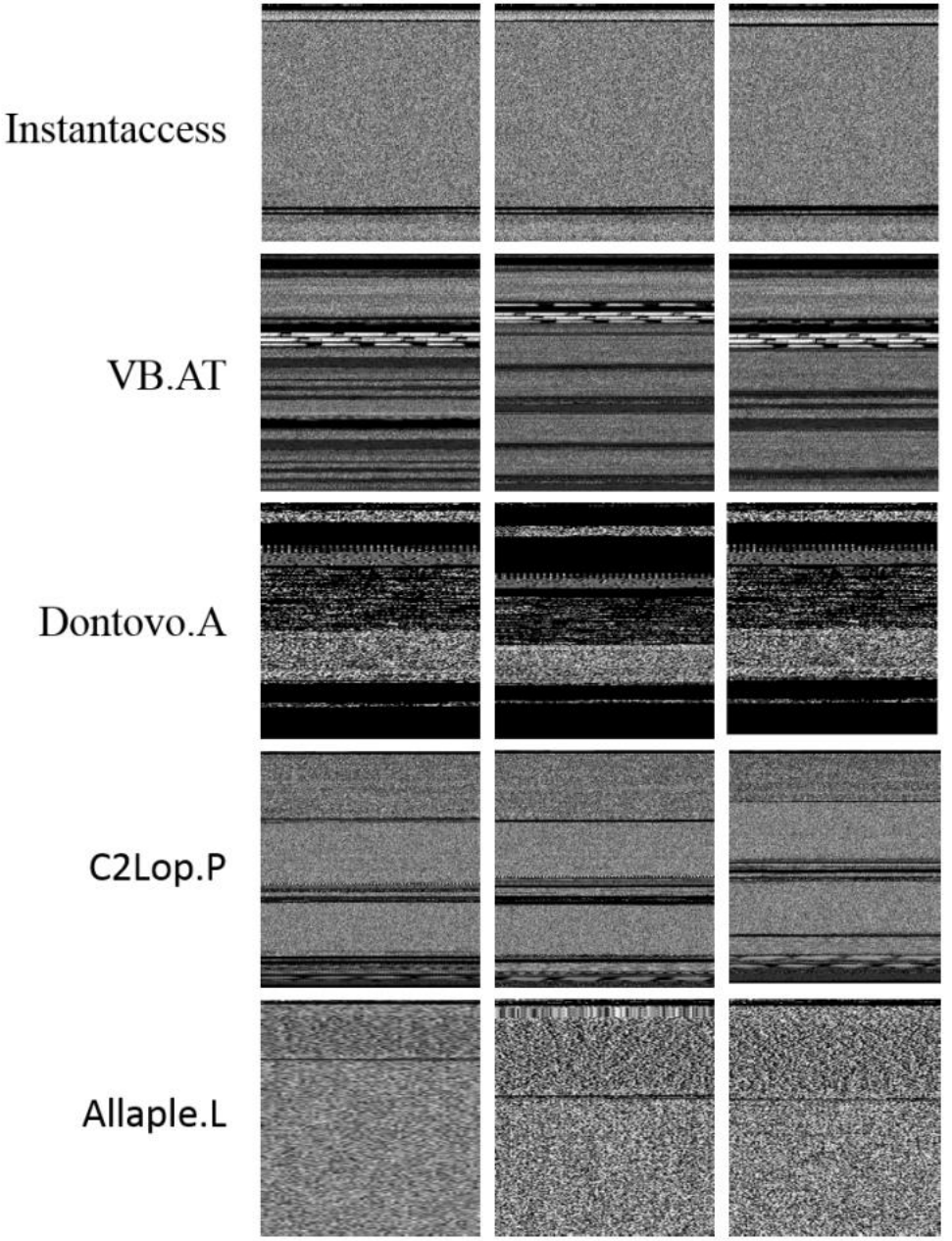


Fig. 2. Malware images of 5 different families. Each row having 3 variants of malware of same family [10]

Some Feature Extraction Methods

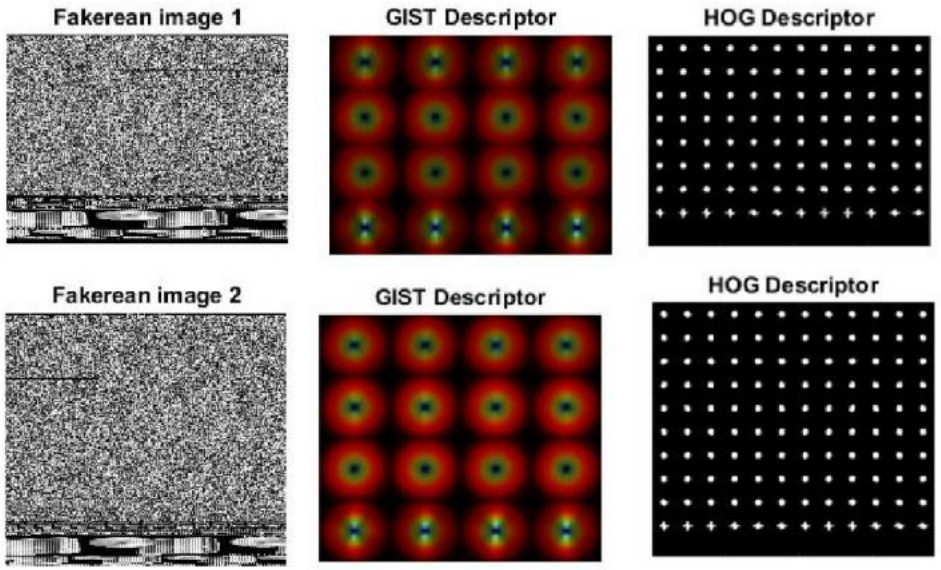


Fig. 3. GIST and HOG descriptor visualization of 2 variants Fakerean malware family.

Upload the result to Moodle

- Upload your code and the ***detailed*** reports to Moodle.
  - **Code**: You must add some comments in your code for easy understanding.
  - **Report Files**: Document (Word) and Presentation (PPT).
    - Including the table outlining team member responsibilities.
  - Zipped to a file. The file name should be “Student ID_HW3.zip”