Feasibility Report

SE 303 Mini Project - Team 12

Problem Statement:

To implement a machine learning based solution for the detection of malicious JPEG images in Python.

Motivation:

Cyber attackers use files sent over the Internet to launch attacks on individuals and organizations. Files such as .exe are considered to be harmful and hence files that are considered harmless by most users such as .pdf, .docx are being increasingly used to hide malicious payloads in them. JPEG is a popular image format. Embedding malicious payloads can be done by exploiting vulnerabilities in JPEG file structure or image viewers.

The JPEG file structure contains elements called markers and these are used by attackers to hide the payloads. So by examining the relative content of these markers in malicious and benign images one can classify the image. This is the basic idea behind MalJPEG.

EED	0 EE	En	00	10	4.0	16	40	16	00	01	01	01	00	40
						-				-	-	-		
100000000000000000000000000000000000000														
09 0	7 0A	0E	0C	0F	0E	0E	0C	0D	0D	0F	11	16	13	0F
10 1	5 11	0D	0D	13	1A	13	15	17	18	19	19	19	0F	12
1B1	B 1B	18	1D	16	18	19	18	FF	DB	00	43	01	04	04
04 0	6 05	06	0B	06	06	0B	18	10	0D	10	18	18	18	18
18 1	8 18	18	18	18	18	18	18	18	18	18	18	18	18	18
18 1	8 18	18	18	18	18	18	18	18	18	18	18	18	18	18
18 1	8 18	18	18	18	18	18	18	18	18	18	18	18	FF	C
00 1	1 08	01	F4	01	E8	03	01	22	00	02	11	01	03	11
01 F	F C4	00	1D	00	00	01	04	03	01	01	00	00	00	00
00 0	0 00	00	00	00	04	02	03	05	06	00	07	08	01	09
FFC	400	40	10	00	02	01	03	03	02	05	03	03	02	05
02 0	5 03	05	01	01	02	03	00	04	11	05	12	21	06	31
071	3 22	41	51	08	61	71	14	32	81	42	91	15	23	52
A1B	1 33	62	16	24	C1	D1	E1	17	72	F1	09	25	43	53
F0 8	2 FF	C4	00	1C	01	00	02	02	03	01	01	00	00	00
00 0	0 00	00	00	00	00	03	04	02	05	01	06	07	00	08
FFC	4 00	39	11	00	02	02	02	01	03	03	02	05	02	05
							70.00							
FFD	9					,								
	00 4 03 0 09 0 10 1 18 1 04 0 18 1 18 1 18 1 00 1 00 0 FFC 02 0 07 1 A1 E F0 8 00 0 FFC	00 48 00 03 03 03 09 07 0A 10 15 11 1B 1B 1B 04 06 05 18 18 18 18 18 18 18 18 18 00 11 08 01 FF C4 00 00 00 FF C4 00 02 05 03 07 13 22 A1 B1 33 F0 82 FF 00 00 00	00 48 00 00 03 03 03 04 09 07 0A 0E 10 15 11 0D 1B 1B 1B 1B 04 06 05 06 18 18 18 18 18 18 18 18 18 18 10 11 08 01 01 FF C400 00 00 00 00 FF C4 00 40 02 05 03 05 07 13 22 41 A1 B1 33 62 F0 82 FF C4 00 00 00 00 FF C4 00 39	00 48 00 00 FF 03 03 03 04 04 09 07 0A 0E 0C 10 15 11 0D 0D 1B 1B 1B 18 1D 04 06 05 06 0B 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 00 11 08 01 F4 01 FF C400 1D 00 00 00 00 FF C4 00 40 10 02 05 03 05 01 07 13 22 41 51 A1 B1 33 62 16 F0 82 FF C4 00 00 00 00 00 FF C4 00 39 11	00 48 00 00 FFDB 03 03 03 04 04 04 09 07 0A 0E 0C 0F 10 15 11 0D 0D13 1B 1B 1B 1B 1B 1D 16 04 06 05 06 0B 06 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 00 11 08 01 F4 01 01 FF C4 00 1D 00 00 00 00 00 00 00 FFC400 40 10 00 02 05 03 05 01 01 07 13 22 41 51 08 A1 B1 33 62 16 24 F0 82 FF C4 00 1C 00 00 00 00 00 00 FFC400 39 11 00	00 48 00 00 FFDB 00 03 03 03 04 04 04 04 05 09 07 0A 0E 0C 0F 0E 10 15 11 0D 0D 13 1A 1B 1B 1B 1B 1B 1D 16 1B 04 06 05 06 0B 06 06 18 00 11 08 01 F4 01 E8 01 FF C4 00 1D 00 00 00 00 00 00 00 00 04 FFC4 00 40 10 00 02 07 13 22 41 51 08 61 A1 B1 33 62 16 24 C1 F0 82 FF C4 00 1C 01 00 00 00 00 00 00 00 FF C4 00 39 11 00 02	00 48 00 00 FFDB 00 43 03 03 03 04 04 04 05 06 09 07 0A 0E 0C 0F 0E 0E 10 15 11 0D 0D 13 1A 13 1B 1B 1B 1B 1D 16 18 19 04 06 05 06 0B 06 06 0B 18 00 11 08 01 F4 01 E8 03 01 FF C4 00 1D 00 00 01 00 00 00 00 00 00 00 02 FFC400 40 10 00 02 01 02 05 03 05 01 01 02 03 07 13 22 41 51 08 61 71 A1 B1 33 62 16 24 C1 D1 F0 82 FF C4 00 1C 01 00 00 00 00 00 00 00 00 03 FF C4 00 39 11 00 02 02	00 48 00 00 FFDB 00 43 00 03 03 03 04 04 04 05 06 0A 09 07 0A 0E 0C 0F 0E 0E 0C 10 15 11 0D 0D13 1A 13 15 1B 1B 1B 1B 1B 1D 16 18 19 18 04 06 05 06 0B 06 06 0B 18 18 18 18 18 18 18 18 18 18 18 18 18	00 48 00 00 FF DB 00 43 00 04 03 03 03 04 04 04 05 06 0A06 09 07 0A 0E 0C 0F 0E 0E 0C 0D 10 15 11 0D 0D 13 1A 13 15 17 1B 1B 1B 1B 1B 1D 16 18 19 18 FF 04 06 05 06 0B 06 06 0B 18 10 18 18 18 18 18 18 18 18 18 18 18 18 18	00 48 00 00 FF DB 00 43 00 04 03 03 03 03 04 04 04 05 06 0A06 06 09 07 0A 0E 0C 0F 0E 0E 0C 0D 0D 10 15 11 0D 0D 13 1A 13 15 17 18 1B 1B 1B 1B 1B 1D 16 18 19 18 FF DB 04 06 05 06 0B 06 06 0B 18 10 0D 18 18 18 18 18 18 18 18 18 18 18 18 18	00 48 00 00 FF DB 00 43 00 04 03 03 03 03 03 04 04 04 05 06 0A06 06 05 09 07 0A 0E 0C 0F 0E 0E 0C 0D 0D 0F 10 15 11 0D 0D 13 1A 13 15 17 18 19 1B 1B 1B 1B 1B 1D 16 18 19 18 FF DB 00 04 06 05 06 0B 06 06 0B 18 10 0D 10 18 18 18 18 18 18 18 18 18 18 18 18 18	00 48 00 00 FF DB 00 43 00 04 03 03 03 03 03 03 03 04 04 04 05 06 0A06 06 05 05 09 07 0A 0E 0C 0F 0E 0E 0C 0D 0D 0F 11 10 15 11 0D 0D 13 1A 13 15 17 18 19 19 18 1B 1B 1B 1B 1D 16 18 19 18 FF DB 00 43 04 06 05 06 0B 06 06 0B 18 10 0D 10 18 18 18 18 18 18 18 18 18 18 18 18 18	00 48 00 00 FF DB 00 43 00 04 03 03 03 03 03 03 03 03 03 03 04 04 04 05 06 0A06 06 05 05 06 09 07 0A 0E 0C 0F 0E 0E 0C 0D 0D 0F 11 16 10 15 11 0D 0D 13 1A 13 15 17 18 19 19 18 1B 1B 1B 1B 1D 16 18 19 18 FF DB 00 43 01 04 06 05 06 0B 06 06 0B 18 10 0D 10 18 18 18 18 18 18 18 18 18 18 18 18 18	

Fig1: JPEG file structure with markers

Implementation:

The steps involved in the MalJPEG solution are as follows.

- 1.MalJPEG receives an image as an input.
- 2.MalJPEG examines the image statically without invoking an image viewer.(Might have a vulnerability)
- 3. Traverse the JPEG file structure to extract the features.
- 4. Transfer the features to a machine learning based model which outputs a classification.

The MalJPEG features are based on the presence and size of specific markers within the JPEG image file structure. This is because malicious images always have something to do with markers. For example, it has been found that some malicious images contain data after the end-of-file (EOI) marker.

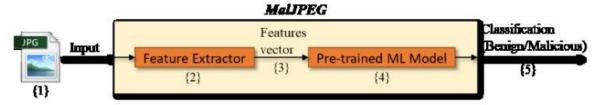


Fig 2: The MalJPEG solution

#	Feature Name	Description	Info Gain Rank
1	Marker_EOI_content after num	Number of bytes after the EOI (end of file) marker.	0.058
2	Marker_DHT_size_max	Maximal DHT marker size found in the file.	0.025
3	File size	Image file size in bytes.	0.023
4	Marker_APP1_size_max	Maximal APP1 marker size found in the file.	0.023
5	Marker_COM_size_max	Maximal COM marker size found in the file.	0.017
6	Marker_DHT_num	Number of DHT markers found in the file.	0.016
7	File_markers_num	Total number of markers found in the file.	0.014
8	Marker_DQT_num	Number of DQT markers found in the file.	0.012
9	Marker_DQT_size_max	Maximal DQT marker size found in the file.	0.012
10	Marker_APP12_size_max	Maximal APP12 marker size found in the file.	0.011

Fig: MalJPEG features.

A total of 10 features are detected . These features have been decided by the researchers of the base paper after closely examining the file structure of malicious and benign images.

How will this help?

Antivirus engines can detect malware only based on its signature database. When a new malware is identified or discovered there is a time lag in updating the signature database. A machine learning solution has no such limitation and can be greatly useful.

Based on the above points, we have decided to implement this solution to the malicious JPEG problem in Python.

Team Members

Akshith Nettar Mahalinga 181IT104 Amith Bhat 181IT103 Laharish S 181IT125