Final Lab Assignment

Date of submission: 05-12-2020

Aim:Detect malware using image visualization approach.

Description: To evade detection, malware authors introduce polymorphism to the malicious components. This means that malicious files belonging to the same malware "family", with the same forms of malicious behavior, are constantly modified and/or obfuscated using various tactics, such that they look like many different files. Dataset comprises 25 malware families with varying number of variants per family. Each malware is represented in the form of grayscale image. The task to be performed as listed below:

Download link for dataset: https://sarvamblog.blogspot.com/2014/08/supervised-classification-with-k-fold.html

[1] Extract features using Local Binary Pattern (LBP)

Refer

(a) https://scikit-image.org/docs/dev/auto_examples/features_detection/plot_local_binary_pattern.html

(b) https://link.springer.com/chapter/10.1007/978-3-540-24670-1_36

[2] Extract texture features using Gray-Level Co-Occurrence Matrix (GLCM)

Refer

(a)
https://support.echoview.com/WebHelp/Windows_and_Dialog_Boxes/Dialog_Boxes/Variable_properties_dialog_box/Operator_pages/GLCM_Texture_Features.htm

(b)Haralick, RM.; Shanmugam, K., “Textural features for image classification” IEEE Transactions on systems, man, and cybernetics 6 (1973): 610-621. DOI:10.1109/TSMC.1973.4309314

(c) Video: https://www.youtube.com/watch?v=XDqlyQ46C7M&list=RDCMUCzwo7UlGkb-8Pr6svxWo-LA&index=1

[*] Create machine learning models using algorithms studied in the course.

[*] Create a 2D-CNN for classifying malware images into their respective families.

[*] Compare classification algorithms based on their performance

[*] Build a stacking model to predict the test samples

Output must contain a confusion matrix, and standard evaluation metrics like accuracy, precision, recall, ROC, etc. Additionally, print the time spent in the training and testing phase.