Malware Detection

Tamir Moshiashvili, Eden Shuker
Guided by: Assaf Barak

CS@BIU
המחלקה למדעי המחשב
אוניברסיטת בר-אילן

אוניברסיטת בר-אילן
Bar-Ilan University

BIU
Center for Research in Applied Cryptography and Cyber Security

Problem

- Classification of benign files and variants of malware files into their respective families.
- Input - file represented as array of bytes together with PE header data.
- Output - class of the file (benign or one of the malware families).

XGBoost Model

Approach

- Represent the file as vector of features.
- Extract features.
- Represents each file as a vector with the extracted features.
- Train xgboost classifier with this vectors.

Malware Family Example - Zbot

This trojan gathers information from your PC and sends to a hacker, which also can get control on your PC, its payload:

- Disable firewall.
- Lowers internet security.
- pass access to hacker.

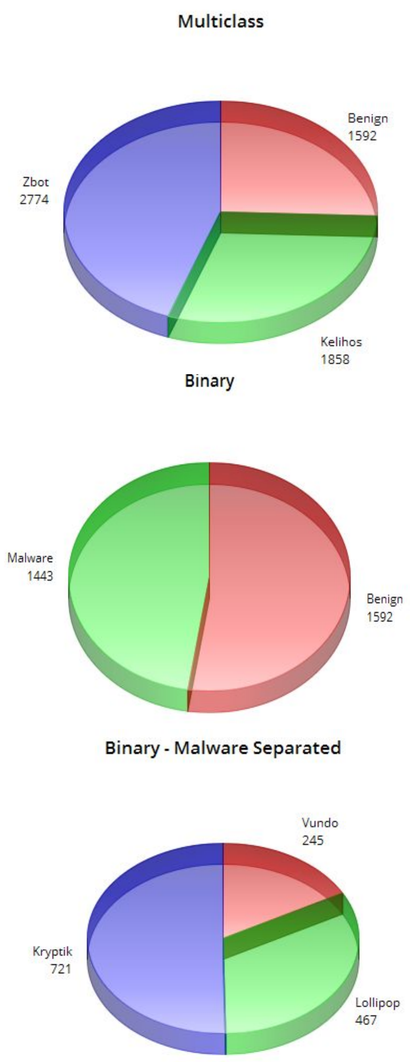
Data

Each file is Windows8 PE **without** the PE header and includes:
.bytes file (hexadecimal representation) and .asm file (disassembly).
Data contains:

- 8 malware classes from the Kaggle contest of 2015.
- One class of bening files from BIU Cyber Center.

Overall 9 classes**.**

Files Distribution



Features Extraction

Ngrams

- For each class - extract 100,000 most common ngrams
- Join ngrams from all classes
- Pick 750 ngrams for each class by using cross-entropy. ngrams array = 750 * # classes
- representing file - vec[i] = 1 if ngrams[i] in file, 0 otherwise

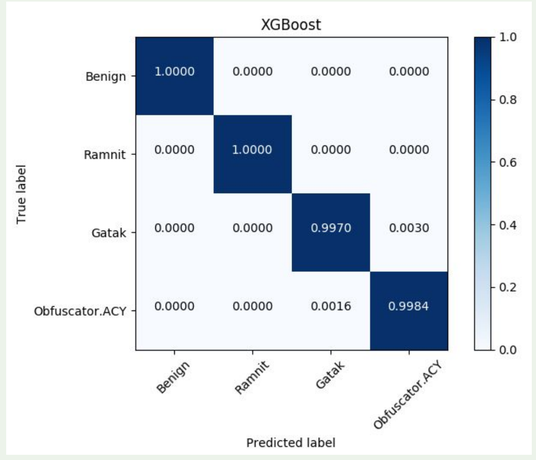
Segments

- For each file available - extract a list of segment names
- Join segment names to create a set
- representing file - vec[i] = # lines in segments[i] in the given file

- Concatenate ngrams vector of a file with its segments vector

Results

Accuracy Scores	
Train	**Test**
99.98	99.62



Deep Learning Model

Approach

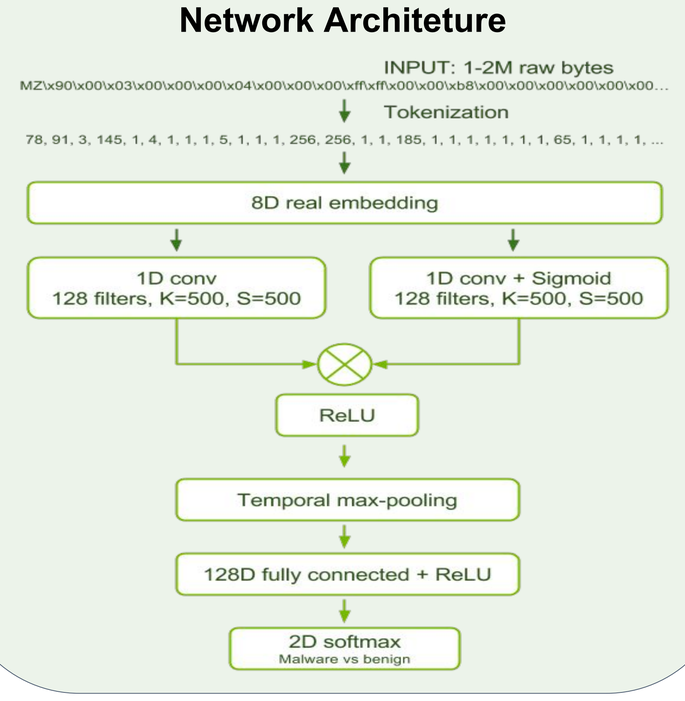
- Represent the file as vector of fixed length (2 million).
- Each value is integer between 0 to 256.
- Pad the vectors for shorter files with zeros.
- Embed Each value in the vector to vector of size 8.
- Train deep-learning model (described below).

Data

Each file is Windows8 PE **with** the PE header. files in format '.bytes' (hexadecimal representation).
All files from BIU Cyber Center.
Data contains:

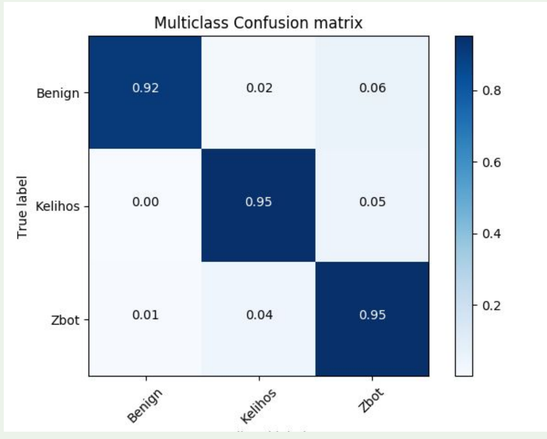
- Multiclass - 2 malwares: Kelihos and Zibot together with Benign files (overall 3 classes).
- Binary - overall 2 classes, Benign and Malware. Malware contains Vundo, Lollipop and Kryptik.

Network Architeture



Results

Accuracy Scores		
Type \ Set	**Train**	**Test**
Binary	99.95	97.18
Multiclass	95.32	90.11



Future Work

Change the bytes of each file without changing its operation and checking the performance of each model on the new files.

python™ dmlc XGBoost PYTORCH Capstone The Ultimate Disassembler radare2