Appendices for: Detection of metamorphic malware packers using multilayered LSTM networks

Erik Bergenholtz[1], Emiliano Casalicchio[1 2], Dragos Ilie[1], and Andrew Moss[1]

[1] Blekinge Institute of Technology {ebz,awm,dil,emc}@bth.se
[2] Sapienza University of Rome, Italy emiliano.casalicchio@uniroma1.it

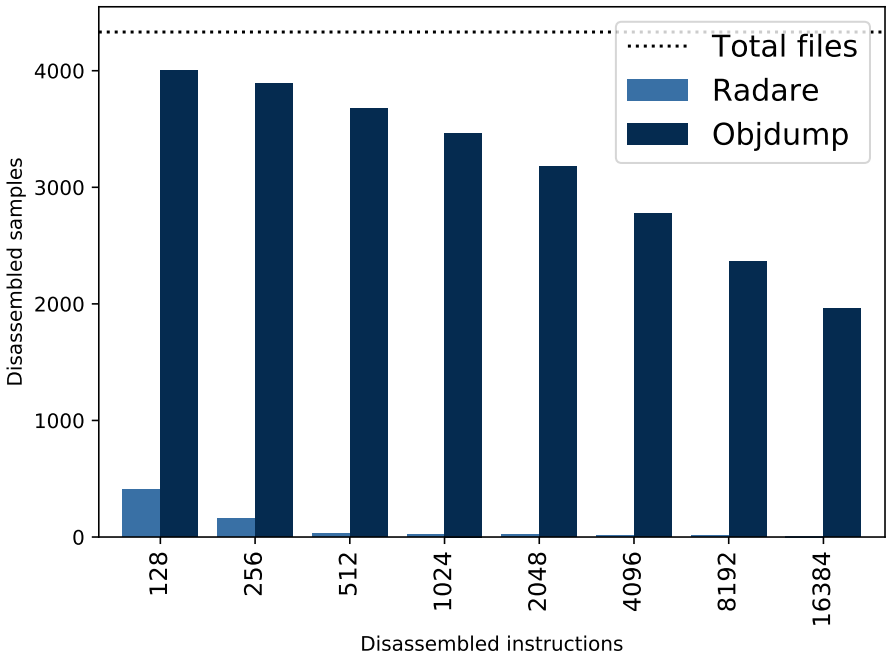
Appendix 1 Data Collection and Filtering

Table 1. Number of files at different stages of the data collection

Original files	2055
Included	1904
Packed executables	63628
Extracted from reference system	78535
Augmented negative set	14906
Final positive	58379
Final positive per packer	1358
Final negative	12549
Final total	70928

Appendix 2 Evaluation of Disassembly Engines

Fig. 1. Comparison between `objdump` and `radare2`



Appendix 3 Selection of Parameters

The selection of the disassembly length and sliding window size were the result of evaluating these parameters on a subset of the full data set. This subset consisted of 100 files from each packer, and all original files, to make sure the data fed to the network was balanced. The evaluation consisted of training one model for each individual packer with the given parameters. The graphs below show the average accuracy and average prediction time for each parameter setting, and the whiskers show the mean square error (MSE) of the values.

Fig. 2. Evaluation of disassembly length

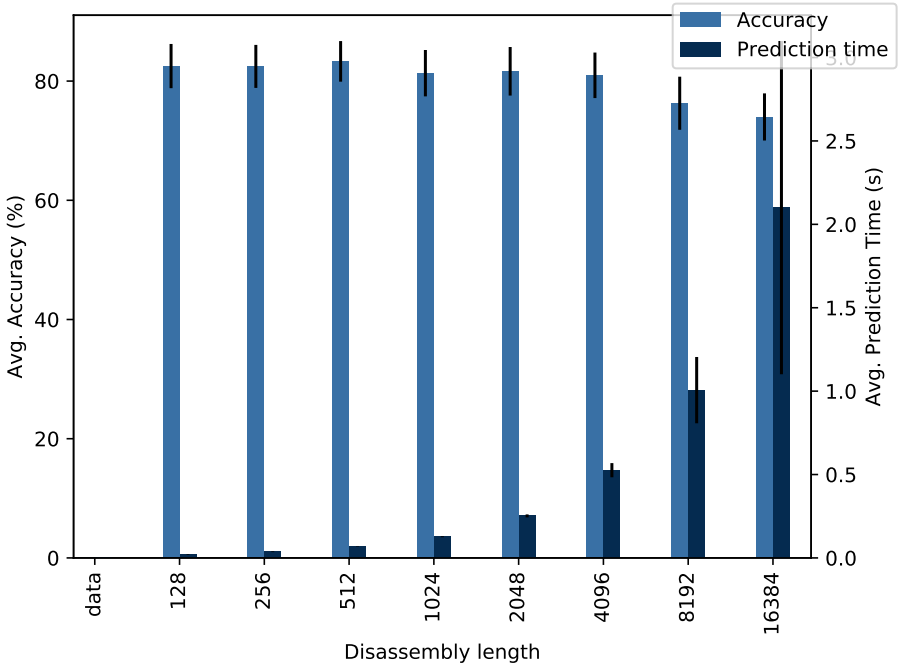


Figure 2 shows the results of evaluating the disassembly length. Eight values were evaluated: $n \in \{128, 256, 512, 1024, 2048, 4096, 8192, 16384\}$. While evaluating these parameters, we used a window size of $w = 1$. The figure shows that the prediction time increases linearly with n, while the accuracy seems generally unchanged. Since high speed is important in a real-time setting, we chose to use a disassembly length of n =128 .

The results of evaluating the window sizes $w \in \{1, 2, 3, 4, 5\}$ is shown in Figure 3. From the graph, it is clear that neither time nor accuracy is greatly affected by this parameter. We chose w =1 because this value had the highest average accuracy, accompanied with the lowest MSE of 1.86.

Fig. 3. Evaluation of sliding window size