Machine Learning and the Malware Arms Race

Abdullah Al-Dujaili, Alex Huang, Erik Hemberg, Una-May O'Reilly

Project Overview

- Malware detectors have been improved and automated with the help of machine learning (ML).
- Unfortunately adversaries aiming to thwart them are also aided by ML (see Fig. 1).
- Our project explores the malware arms race of artificially intelligent competitors for ways detectors can gain the upper hand.

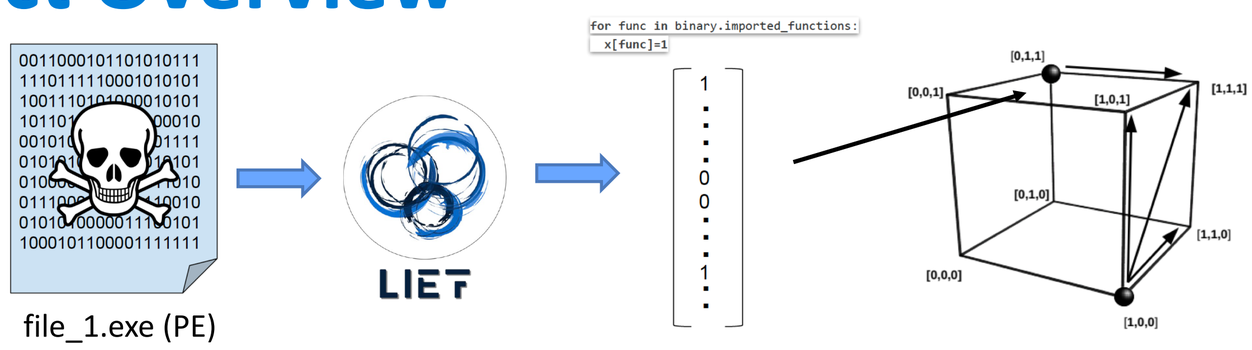

Fig. 1. An appropriate feature vector of a PE can be built based on the PE's imported functions. A malware author can import additional functions without affecting the malicious functionality giving rise to a set of adversarial versions of the same PE. These adversarial versions can be found with the help of an ML algorithm.

Current Progress

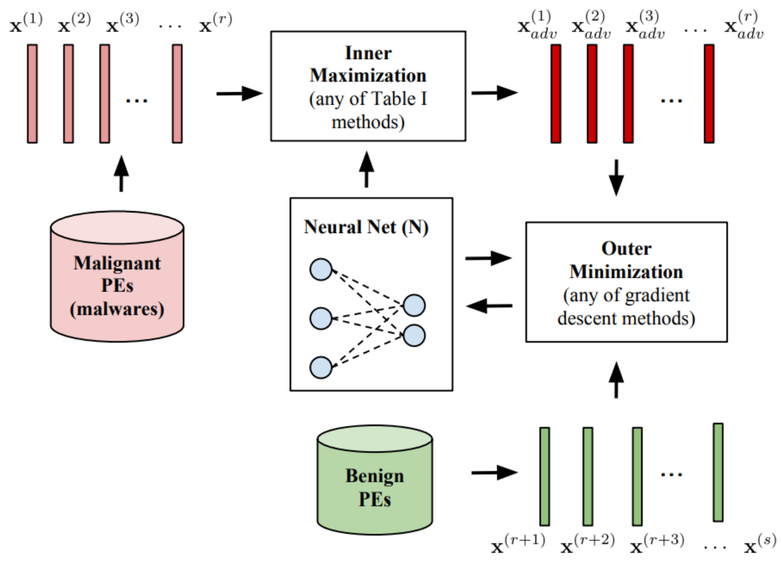

Fig. 2. SLEIPNIR: a saddle-point formulation for hardening ML models with binary feature space.

Table I. Performance Metrics

Model	Accuracy	FPR	FNR	$\bar{\mathcal{N}}_{\text{BS}}$
Natural	91.9	8.2	8.1	1.0
dFGSM^k	+0.1	+1.4	−1.7	+1.6
rFGSM^k	−0.6	+3.6	**−2.4**	**+3.0**
BGA^k	**+0.2**	+0.0	−0.5	+2.5
BCA^k	−0.3	+0.9	−0.5	+0.0
[Grosse et al.'17]	−1.1	**−3.9**	+5.9	+0.6

Table II. Evasion Rates

Model	Adversary					
	Natural	dFGSM^k	rFGSM^k	BGA^k	BCA^k	[Grosse et al.'17]
Natural	8.1	99.7	99.7	99.7	41.7	99.7
dFGSM^k	6.4	**6.4**	21.1	7.3	27.4	99.2
rFGSM^k	**5.7**	7.0	**5.9**	**5.9**	**6.8**	35.0
BGA^k	7.6	39.6	17.8	**7.6**	10.9	68.4
BCA^k	7.6	99.5	99.5	91.8	**7.9**	98.6
[Grosse et al.'17]	14.0	69.3	69.3	37.5	14.1	**15.6**

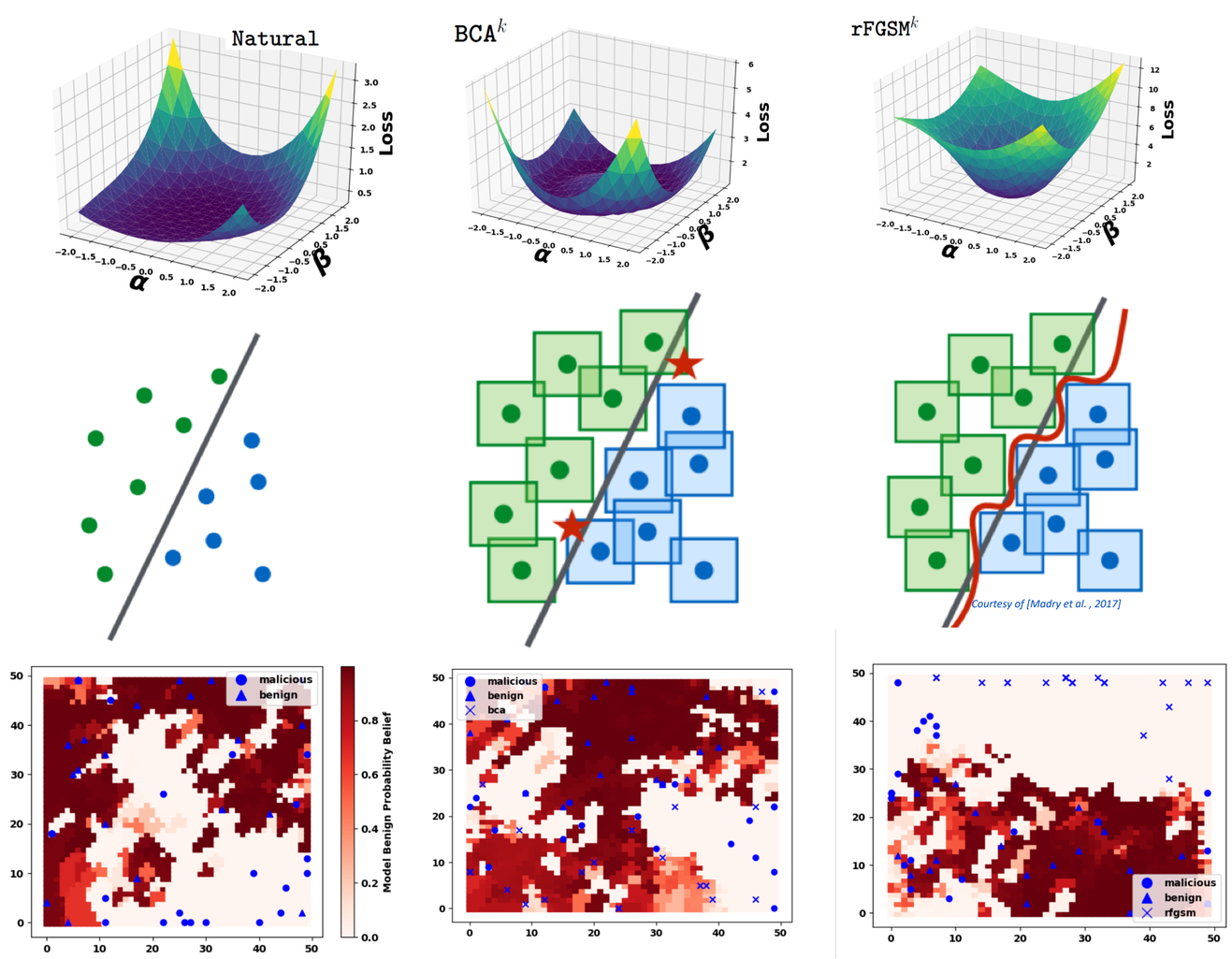

Fig. 3. In contrast to standard generalization, the decision landscape (input space) has a stronger association with the hardened model's robust generalization, compared to the geometry of the loss landscape (parameter space).

Relevant Papers

1. Al-Dujaili, et al. "Adversarial Deep Learning for Robust Detection of Binary Encoded Malware." IEEE SPW, 2018.
2. Huang et al. "On Visual Hallmarks of Robustness to Adversarial Malware." IJCAI-IREDLIA, 2018..

Open Questions & Future Vision

- Static Analysis vs. Dynamic Analysis
- Representation Learning: Abstract Syntax Trees (ASTs)
- Interpreted languages: PowerShell Scripts

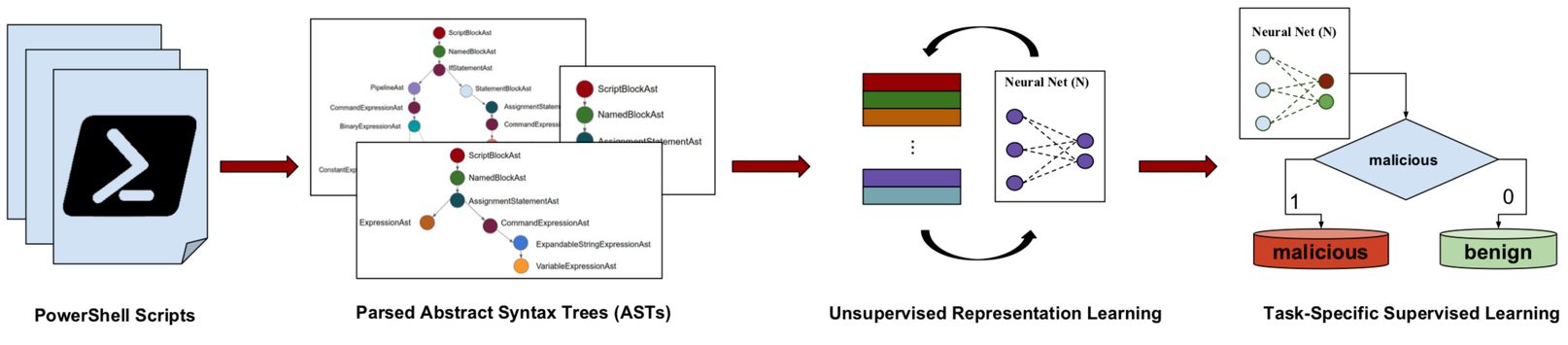

Fig. 4. AST-based deep learning for malicious PowerShell detection.

#AI Research Week

hosted by MIT-IBM Watson AI Lab

Detailed Contact Information
(corresponding author): Abdullah Al-Dujaili, aldujail@mit.edu