

Contents

Part I The Beginnings of Adversarial ML

1	Introduction	3
1.1	Motivation	3
1.2	Problem Statement	4
1.3	Organization of Dissertation	7
2	Background	9
2.1	Malware	9
2.2	Portable Executable Format	10
2.3	Related Work	12
2.3.1	Gradient Optimization	13
2.3.2	Generative Adversarial Networks	14
2.3.3	Reinforcement Learning	15
2.3.4	Genetic Programming	16
2.3.5	Universal Adversarial Perturbations	18
2.4	Summary	19

Part II Framework for Adversarial Malware Evaluation

3	FAME	23
3.1	Notation	23
3.1.1	Feature-Space Attacks	24
3.1.2	Feature-Space Constraints	24
3.1.3	Problem-Space Attacks	24
3.1.4	Problem-Space Constraints	24
3.2	Threat Model	25

3.2.1	Adversary Objectives	25
3.2.2	Adversary Knowledge	25
3.2.3	Adversary Capabilities	26
3.3	Experimental Settings	26
3.3.1	Target Model	27
3.3.2	Datasets	27
3.3.3	Byte-Level Perturbations	29
3.3.4	Integrity Verification	30
3.4	Summary	31

Part III Problem-Space Attacks

4	Stochastic Method	35
4.1	Requirements	36
4.2	Methodology	36
4.2.1	Toolbox	37
4.2.2	Verifier	37
4.2.3	Classifier	37
4.3	Evaluation	38
4.3.1	Integrity	38
4.3.2	Evasiveness	40
4.3.3	High-Dimensional Sequences	41
4.4	Summary	42
5	Genetic Programming	43
5.1	Requirements	44
5.2	Methodology	45
5.2.1	Genetic Operators	46
5.2.2	Fitness Function	47
5.3	Evaluation	47
5.4	Cross-evasion	49
5.5	Summary	50
6	Reinforcement Learning	51
6.1	Methodology	52
6.1.1	State	53
6.1.2	Action	53
6.1.3	Reward	53
6.1.4	Penalty	55
6.1.5	Model Parameters	55

6.2	Evaluation	56
6.2.1	Reset	56
6.2.2	Diversity	56
6.2.3	Evasion Rate	57
6.3	Summary	60
7	Universal Attacks	61
7.1	Requirements	62
7.2	Universal Evasion Rate	62
7.3	UAP Search	62
7.4	Evaluation	64
7.5	Summary	65
 Part IV Feature-Space Attacks		
8	Gradient Optimization	69
8.1	Convolutional Neural Networks	69
8.2	Methodology	70
8.3	Summary	72
9	Generative Adversarial Nets	73
9.1	GANs in Security	74
9.2	Methodology	74
9.3	Summary	76
 Part V Benchmark & Defenses		
10	Comparison of Strategies	79
10.1	Benchmark Settings	79
10.2	Summary	81
11	Towards Robustness	83
11.1	Feature Reduction	84
11.2	UAP-Based Adversarial Training	85
11.2.1	Evaluate Hardened Models	87
11.2.2	UAP Search Alternative	89
11.3	Summary	91

Part VI Closing Remarks

12 Conclusions & Outlook	95
12.1 Revising Research Questions	96
12.2 Contributions	99
12.2.1 Empirical Guidance for Malware Attacks	99
12.2.2 Adversarial Defenses derived from UAP attacks	100
12.2.3 Framework to Evaluate Malware Classifiers	100
12.3 Connecting the dots	101
12.4 Future Work	102
List of Publications	105
References	107