

Contents

Biography

xi

Preface

xiii

PART 1 Preliminaries

1. Background and motivation	3
1.1. What is adversarial machine learning?	3
1.2. Mathematical notations	4
1.3. Machine learning basics	5
1.4. Motivating examples	7
1.5. Practical examples of AI vulnerabilities	12
1.6. Open-source Python libraries for adversarial robustness	12

PART 2 Adversarial attack

2. White-box adversarial attacks	15
2.1. Attack procedure and notations	16
2.2. Formulating attack as constrained optimization	17
2.3. Steepest descent, FGSM and PGD attack	19
2.4. Transforming to an unconstrained optimization problem	20
2.5. Another way to define attack objective	21
2.6. Attacks with different ℓ_p norms	22
2.7. Universal attack	22
2.8. Adaptive white-box attack	23
2.9. Empirical comparison	24
2.10. Extended reading	26
3. Black-box adversarial attacks	29
3.1. Evasion attack taxonomy	29
3.2. Soft-label black-box attack	30
3.3. Hard-label black-box attack	34
3.4. Transfer attack	38
3.5. Attack dimension reduction	39
3.6. Empirical comparisons	40
3.7. Proof of Theorem 1	43
3.8. Extended reading	45

4. Physical adversarial attacks	47
4.1. Physical adversarial attack formulation	47
4.2. Examples of physical adversarial attacks	48
4.3. Empirical comparison	50
4.4. Extending reading	50
5. Training-time adversarial attacks	51
5.1. Poisoning attack	51
5.2. Backdoor attack	52
5.3. Empirical comparison	53
5.4. Case study: distributed backdoor attacks on federated learning	53
5.5. Extended reading	57
6. Adversarial attacks beyond image classification	59
6.1. Data modality and task objectives	59
6.2. Audio adversarial example	59
6.3. Feature identification	60
6.4. Graph neural network	60
6.5. Natural language processing	61
6.6. Deep reinforcement learning	66
6.7. Image captioning	66
6.8. Weight perturbation	67
6.9. Extended reading	69
PART 3 Robustness verification	
7. Overview of neural network verification	73
7.1. Robustness verification versus adversarial attack	73
7.2. Formulations of robustness verification	75
7.3. Applications of neural network verification	76
7.4. Extended reading	77
8. Incomplete neural network verification	79
8.1. A convex relaxation framework	79
8.2. Linear bound propagation methods	80
8.3. Convex relaxation in the dual space	85
8.4. Recent progresses in linear relaxation-based methods	85
8.5. Extended reading	87

9. Complete neural network verification	89
9.1. Mixed integer programming	89
9.2. Branch and bound	90
9.3. Branch-and-bound with linear bound propagation	92
9.4. Empirical comparison	93
10. Verification against semantic perturbations	95
10.1. Semantic adversarial example	95
10.2. Semantic perturbation layer	97
10.3. Input space refinement for semantify-NN	102
10.4. Empirical comparison	105
PART 4 Adversarial defense	
11. Overview of adversarial defense	113
11.1. Empirical defense versus certified defense	113
11.2. Overview of empirical defenses	114
12. Adversarial training	119
12.1. Formulating adversarial training as bilevel optimization	119
12.2. Faster adversarial training	121
12.3. Improvements on adversarial training	122
12.4. Extended reading	125
13. Randomization-based defense	127
13.1. Earlier attempts and the EoT attack	127
13.2. Adding randomness to each layer	128
13.3. Certified defense with randomized smoothing	132
13.4. Extended reading	136
14. Certified robustness training	137
14.1. A framework for certified robust training	137
14.2. Existing algorithms and their performances	139
14.3. Empirical comparison	141
14.4. Extended reading	142
15. Adversary detection	143
15.1. Detecting adversarial inputs	143
15.2. Detecting adversarial audio inputs	146

15.3. Detecting Trojan models	149
15.4. Extended reading	155
16. Adversarial robustness of beyond neural network models	157
16.1. Evaluating the robustness of K-nearest-neighbor models	158
16.2. Defenses with nearest-neighbor classifiers	166
16.3. Evaluating the robustness of decision tree ensembles	170
17. Adversarial robustness in meta-learning and contrastive learning	183
17.1. Fast adversarial robustness adaptation in model-agnostic meta-learning	183
17.2. Adversarial robustness preservation for contrastive learning: from pretraining to finetuning	188
PART 5 Applications beyond attack and defense	
18. Model reprogramming	201
18.1. Reprogramming voice models for time series classification	201
18.2. Reprogramming general image models for medical image classification	206
18.3. Theoretical justification of model reprogramming	209
18.4. Proofs	212
18.5. Extended reading	215
19. Contrastive explanations	217
19.1. Contrastive explanations method	217
19.2. Contrastive explanations with monotonic attribute functions	220
19.3. Empirical comparison	223
19.4. Extended reading	225
20. Model watermarking and fingerprinting	227
20.1. Model watermarking	227
20.2. Model fingerprinting	232
20.3. Empirical comparison	236
20.4. Extended reading	239
21. Data augmentation for unsupervised machine learning	241
21.1. Adversarial examples for unsupervised machine learning models	241
21.2. Empirical comparison	245
<i>References</i>	251
<i>Index</i>	273