

Index

A

Accuracy

- certified robust, 114
- classification, 10, 195, 248
- clean, 114, 155
- drop, 232
- increase, 250
- prediction, 7, 8, 128
- reprogramming, 204
- standard, 8, 187, 189, 195

Acoustic model (AM), 202–204, 206

Adaptive

- antiwatermark schemes, 229
- attacks, 24, 146

Advanced

- attacks, 146
- finetuning techniques, 189

AdvCL, 189, 191, 194, 197, 198

AdvCL framework, 191

Adversarial

- agent models, 66
- attack, 15–18, 29, 48, 59, 62, 64, 73, 95, 138, 143, 146, 148, 185, 234
- attack formulation, 47
- attacking algorithms, 241
- audio, 148
 - attacks, 149
 - examples, 155
 - inputs, 146
- autoencoder, 247
- characteristics, 148
- context, 189, 193
- contrastive
 - pretraining framework, 189
 - training, 195
- detection, 143, 144, 146
- example, 8, 15–20, 30, 31, 35, 38–40, 47, 48, 60, 62–64, 73, 74, 113–117, 119, 120, 127, 143–145, 159, 186, 188, 191, 232, 241, 243
- detection, 143

for unsupervised machine learning, 241

from natural examples, 144

transferability, 38, 235

inputs, 147

instances, 148

loss, 180

network, 97

perturbation, 15, 22, 23, 41, 43, 66, 73, 74, 116, 127, 147, 153, 158, 160, 162, 164, 193, 194

purposes, 201

robustness, 7–10, 12, 24, 59, 68, 73, 76, 113, 115, 130, 132, 170, 183, 184, 186, 188, 189, 250

for machine learning, 7

in machine learning algorithms, 3

in MAML, 183, 184

targets, 148

threats, 3

timing attacks, 66

unsupervised example, 244

view, 193

Adversarial full finetuning (AFF), 189–191, 195, 197

Adversarial linear finetuning (ALF), 189, 191, 198

Adversarial machine learning (AdvML), 3, 201

Adversarial reprogramming (AR), 201, 206

Adversarial training (AT), 10, 114, 115, 117, 119, 121, 122, 128, 131, 180, 185, 189–191

algorithm, 20

fast, 121

method, 107, 129

process, 120

Adversarially robust metric learning (ARML), 168, 170

algorithm, 170

certified robustness, 170

Adversary, 228, 229

Aggregated prediction, 204

Alignment loss, 211, 212

Antiwatermark
 attacks, 228
 schemes adaptive, 229

Area under curve (AUC), 148, 149, 154, 155

Attack
 efficiency, 41
 fail, 24
 formulation unsupervised, 244, 245
 function for targeted attack, 244
 function unsupervised, 244
 generation, 186, 245
 goal, 17
 loss, 47
 loss function, 59, 63
 methods, 114
 objective, 21, 31, 52
 performance, 22, 63, 113
 problem, 19
 procedure, 16
 process, 47
 results, 41
 scenarios, 143
 semantic, 98, 99, 101, 107
 space, 103
 success, 245
 criterion, 247
 evaluation function, 244
 successfulness, 17, 18, 21, 24
 supervised, 245
 targeted, 16–18, 22, 24, 31, 35, 107
 threat model, 99
 Trojan, 52, 149, 153
 unsupervised, 244

Attack accuracy (AA), 195–197

Attack success rate (ASR), 10, 21, 24, 41, 50, 52, 57

Attacker, 15, 23, 29, 52, 115, 117
 knowledge, 29
 objective function, 52

Attacking
 algorithms adversarial, 241
 methodology, 56
 scheme, 54

Audio
 data inputs, 146
 inputs, 146

Audio adversarial
 attacks, 146
 examples, 59, 69, 146, 147
 perturbation, 60

Augmented
 dataset, 247
 poisoned dataset, 52

Autism Spectrum Disorder (ASD)
 classification, 207
 dataset, 208
 task, 208, 209

Autoencoder (AE), 40, 116, 218, 244, 247
 adversarial, 247
 sparse, 247

Automatic speech recognition (ASR), 146, 147

AutoZOOM
 attack, 41
 method, 39

B

Backdoor
 attacks, 51–54, 149
 in TrojanNets, 150
 process, 154
 trigger, 56

Backdoored model, 51–53, 150

Base
 classifier, 132, 133, 174
 model, 232, 233, 236, 239
 model features, 234

Bayesian neural network (BNN), 131

Benchmarking dataset, 18

Bilevel optimization problem, 184, 186

Binary
 classification, 176
 classification problem, 134
 classifiers, 143, 221
 prediction, 133

Bit error rate (BER), 232

Black-box attack, 29

Box adversarial reprogramming (BAR), 206, 207

Branch-and-Bound (BaB) method, 90

C

Carrier nodes, 230–232

Centralized

attack, 54, 56

backdoor attack, 54, 56

Certifiable

robustness bound, 161

Certificated robustness, 69

Certified

defense, 113, 114, 132, 138, 141, 179

methods, 77, 114

network, 165

robust

accuracy, 114

error, 138, 165, 167, 168

loss, 137

radius, 135, 137

training, 137, 139, 141

training error, 166

tree ensemble, 181

robustness, 10, 134, 135, 137, 140, 165

accuracy, 142

bound, 134

guarantees, 137

training, 139, 141, 142, 180, 181

Character error rate (CER), 148

CIFAR models, 107

Clarifai

moderation API, 208

reprogramming, 208

Class

labels, 173, 208

prediction, 30, 31, 34, 203, 204, 218

prediction results, 29

prediction score, 18

target, 17, 18, 22, 49, 144, 150, 152, 153, 206

Classification

accuracy, 10, 195, 248

benchmark, 204

error, 8

function, 31

loss, 63, 115

model, 73, 132

performance, 170

rules, 38

tasks, 10, 51, 202

tree, 173

Classifier

deep, 117

linear, 123, 188, 190, 195

neural network, 100, 209–211

results, 225

robust, 115

Clean

accuracy, 114, 155

data, 53, 123, 143, 150, 153

dataset, 155

Cluster labels, 195

Color image dataset, 246

Complete verification, 89, 90, 92

methods, 93

problem, 89

Connectionist temporal classification

(CTC), 60

Contrast perturbation, 101

Contrastive

explanations, 217, 220

loss, 191, 193, 248

training adversarial, 195

Contrastive explanations method (CEM),

217, 225

Contrastive explanations method using

monotonic attribute functions

(CEM-MAF), 221, 222

Contrastive learning (CL), 183, 188, 194,

248

Convex

loss, 39

loss functions, 33

relaxation, 61, 80, 84–86

relaxation barrier, 85, 86

relaxation framework, 79, 84

Convolution

layer, 243

layer output, 243

Convolutional autoencoder (CAE), 223,

247

Convolutional neural network (CNN), 10,

41, 63, 66, 105, 107, 129

Coordinatewise

gradient estimation, 31, 32, 41

neuron activation, 154

Crafting

- adversarial attacks, 23
- adversarial examples, 16, 22, 66
- perturbations, 193
- physical adversarial examples, 48

Cross entropy (CE) loss, 17, 52

CROWN, 83, 86, 93, 139

- bound, 83
- bound propagation, 141
- training, 141

D

Data

- augmentation, 107, 186, 241, 242, 246–248
- clean, 53, 123, 143, 150, 153
- input, 3, 6, 15, 29, 51, 52, 67, 208, 217, 243
- input reprogramming, 204
- reconstruction, 244, 247, 248
- samples, 4–8, 30, 31, 40, 52, 53, 96, 149, 186, 207, 210, 242–244
- unlabeled, 188, 194

Datasets, 7, 18, 60, 142, 165, 170, 194, 197, 202, 207, 221, 228, 243, 246, 248

- finetuning, 189
- for attack performance evaluation, 52
- pretraining, 189

Deep

- classifier, 117
- learning models, 201, 227
- pretrained acoustic classification model, 203

Deep neural network (DNN), 6, 10, 22, 89, 116, 206, 227

- models, 227, 233
- robustness, 113, 127

Deep reinforcement learning (DRL), 66

Defense

- approaches, 158
- certified, 113, 114, 132, 138, 141, 179
- components, 118
- framework, 117
- ineffective, 24
- mechanism, 24, 114
- methods, 113–115, 128, 185

methods certified, 77, 114

- model, 113
- network certified, 165
- performance, 113

Deteriorated robustness, 129

Digital watermarking, 227

Dimensionality constraint, 107

Discrete

- inputs, 62
- layers, 130
- models, 157
- models robustness, 158
- nonneural network models, 157
- perturbation setting, 77

Discrete cosine transform (DCT), 235

Discretely parameterized perturbations, 98

Distributed backdoor attack (DBA), 54, 56

Downstream tasks, 183, 187–189, 191, 241

E

EAD attack, 22, 24

Electrocardiogram (ECG), 201

- classes, 204
- classification, 204

Encoder, 40, 97, 116, 117, 247

- for data reconstruction, 248
- part, 117

Ensemble

- feature attribution, 146
- stump verification, 175

Evasion attack, 15, 29, 113

Evasion attack taxonomy, 29

Eventual learning objective AdvCL, 194

Expectation over transformation (EOT), 47, 127

- attack, 48, 127–129
- attack performance, 50

Exponential loss, 167

F

Fashion MNIST, 170, 246

Fast gradient sign method (FGSM), 19, 20, 23, 113, 115, 128, 188

Feature attribution, 145

- maps, 145
- methods, 144
- values, 146

Federated learning (FL), 53–55, 57
 Finetuning, 188–190, 195, 197, 207
 datasets, 189
 efficiency, 189
 performance, 207
 supervised, 189, 190
 Fingerprint transferability, 235
 Formal verification community, 83
 Formulating
 adversarial training, 119
 attack, 17

G

Gaussian augmentation (GA), 247, 248
 Generative adversarial network (GAN), 3, 116, 221
 Generative models, 96
 German Traffic Sign Benchmark (GTSRB)
 dataset, 107, 209
 models, 107
 Gradient
 estimate, 32, 33
 estimation, 32, 37, 39, 41, 128
 estimator, 32–34, 44
 loss function, 121
 regularization, 64
 Gradient boosting decision tree (GBDT), 157, 158, 170, 176
 Gradient descent (GD), 5, 34, 37, 39, 184
 Gradient mean (GM), 232, 234
 GradSigns, 227, 229–231
 Graph Neural Network (GNN), 60, 61, 93
 Groundtruth class label, 4

H

Hidden
 layer, 101, 247
 layer output, 128
 neurons, 149
 High-frequency component (HFC), 193, 194
 Hinge loss, 21, 167
 HSL
 color space, 101
 space, 100

 space attacks, 107
 space perturbation, 107, 108
 Human imperceptible perturbation, 15

I

ImageNet, 7, 8, 24, 26, 41, 127, 154, 165, 194, 206
 competition yearly, 7
 containing, 7
 dataset, 7, 32, 41, 236, 238
 models, 7, 8
 models pretrained, 201
 Inactive neurons, 140
 Incomplete verification solvers, 92
 Incorrect predictions, 49
 Independent verification problem, 90
 Inputs
 adversarial, 147
 discrete, 62
 gradient, 229, 230
 Integrated gradient (IG), 145
 Intellectual property (IP), 227
 infringements, 227, 228
 protection, 235
 protection methods, 232
 Interval bound propagation (IBP), 139, 140
 training, 140, 141
 Inverse DCT (IDCT), 235
 Iterative FGSM (I-FGSM) attack, 20

K

K-nearest neighbor (KNN), 157, 165, 170
 classifiers, 157
 models certified robustness, 170
 models robustness, 168

L

Label
 change, 152
 flipping, 51
 mapping, 209, 210
 prediction, 30, 152
 target, 53, 150, 152, 153, 203
 Layerwise
 convex relaxation, 83
 nonlinear activations, 103

- Leave-one-out (LOO)
 - feature attribution method, 144
 - method designs, 144
- Lightweight
 - finetuning scheme, 191
 - standard linear finetuning, 189
- Linear
 - classifier, 123, 188, 190, 195
 - finetuning
 - settings, 198
 - strategies, 197
 - types, 197
 - prediction head, 191
 - relaxation, 86, 102, 103, 105, 139
- Linear programming (LP), 91, 92
- Local
 - attackers, 56
 - models, 56
- Local intrinsic dimension (LID), 144
- Logistic loss, 167
- Longest common prefix (LCP), 148
- Loss
 - adversarial, 180
 - attack, 47
 - certified robust, 137
 - classification, 63, 115
 - contrastive, 191, 193, 248
 - function, 5, 6, 18, 22, 48, 52, 60, 65, 121, 123, 137, 138, 168, 218, 234
 - function gradient, 121
 - landscapes for gradient estimation, 34
 - MAML, 185
 - objective reprogramming, 210
 - prediction, 185
 - term, 151, 154, 245
- M**
- Machine learning (ML), 5, 6, 8, 29, 149, 183, 201, 227
 - accelerators, 83
 - adversarial, 3, 201
 - algorithms, 3
 - basics, 5
 - interpretability, 77
 - models, 3, 6, 9, 15, 16, 22, 38, 53, 59, 113, 157, 158, 201, 227, 241, 243
 - paradigms, 183
 - supervised, 6
 - system, 3
 - tasks, 6, 244
 - tools, 29
 - unsupervised, 6
- Manipulated datasets, 51
- Median absolute deviation (MAD), 153
- Minimum
 - adversarial perturbation, 74, 161
 - perturbation norm, 159
- MinMax
 - algorithm, 243, 245
 - attack, 48, 50
 - attack algorithm, 245, 247
 - attack problem, 245
 - optimization problem, 245
- Misclassification, 166
- Misclassification rate, 97
- Mixed integer programming (MIP)
 - problem, 89
 - solver, 93
- MNIST, 24, 26, 41, 103, 107, 141, 142, 170, 201, 223, 246, 247
 - dataset, 41, 142, 165
- Model
 - discrete, 157
 - for medical image classification, 206
 - for reprogramming, 204
 - ImageNet, 7, 8
 - machine learning, 3, 6, 9, 15, 16, 22, 38, 53, 59, 113, 157, 158, 201, 227, 241, 243
 - parameters, 5, 15, 29, 51, 52, 141, 183, 184, 209, 229
 - performance, 8, 68
 - prediction, 4, 8, 15, 29, 144, 176, 208, 210, 218–220
 - reprogramming, 201, 203, 206, 209
 - supervised, 242
 - vendor, 228, 230
 - watermarking, 227
- Model-agnostic meta-learning (MAML), 183–187
 - framework, 184
 - loss, 185

Multiclass classification
 models, 135
 problem, 16, 135
 Multilayer perceptron (MLP), 105, 107
 Multiview CL loss, 191
 Mutual information (MI), 242, 243
 Mutual information neural estimator
 (MINE), 242, 243

N

Natural examples, 116, 144–146, 218
 Natural language processing (NLP) tasks, 10
 Nature language processing (NLP)
 attacks, 62, 63
 models, 62
 Nearest-neighbor (NN)
 classifier, 158, 163, 165, 167
 model, 158–162, 165, 170
 robustness, 163
 Neural Cleanse (NC), 154, 155
 Neural networks, 6, 7, 16, 38, 63, 64, 67,
 68, 73, 76, 77, 82, 95, 97, 100, 113,
 116, 117, 121, 127, 129, 130, 137,
 138, 149, 157, 158, 165, 217, 218,
 232, 242
 adversarial robustness, 6, 157
 classifier, 100, 209–211
 deep, 6, 10, 22, 89, 206, 227
 function, 98
 implementations, 33
 learning, 180
 model, 99, 115, 117, 137, 150, 154, 213,
 217, 243
 parameters, 234
 prediction, 132
 robustness, 68, 69, 115
 robustness models, 113
 training, 17
 verification, 76, 77, 138
 algorithm, 139
 methods, 77, 79, 137
 technique, 138
 Neuron
 activation, 153
 coordinates, 153, 154
 ReLU, 90
 representation, 153

Nonconvex attack spaces, 103
 Nonreprogrammable entry, 203
 Nonrobust loss, 180
 Not Safe For Work (NSFW), 208
 Numerical gradient, 33

O

Obstruct
 verification, 229
 watermark verification, 229
 Occlusion, 98, 99
 attack, 99
 patch, 99
 Opt attack, 41, 148
 Optimal
 attack value, 106
 perturbation, 160, 175
 Optimal transport (OT), 212
 Optimization problem, 18, 19, 23, 31, 35,
 37, 47, 60, 62, 79, 119, 148, 151,
 157, 185, 217–219, 221
 unconstrained, 20

P

Paired perturbation view, 193
 Paraphrasing attack, 63, 64
 Performance
 attack, 22, 63, 113
 benchmarking, 3
 classification, 170
 defense, 113
 degradation, 232, 248
 evaluation, 16, 30
 finetuning, 207
 models, 8, 68
 verification, 102
 Pertinent negative (PN), 217–219, 223
 Pertinent positive (PP), 217, 218, 223
 Perturbation
 adversarial, 15, 22, 23, 41, 43, 66, 73,
 74, 116, 127, 147, 153, 158, 160,
 162, 164, 193, 194
 magnitude, 105
 range, 244
 sensitivity, 68
 set, 62, 63, 120, 137
 space, 63

- techniques, 95
- Trojan, 150
- universal, 22, 150, 152
- Perturbed
 - data sample, 244
 - example, 16, 185
- Pixelwise perturbations, 150
- Poisoned
 - dataset, 51
 - training data, 150
 - training dataset, 51
- Poisoning attack, 51, 52
- Pooling layer, 247
- Prediction
 - accuracy, 7, 8, 128
 - APIs, 206–208, 228
 - binary, 133
 - class, 30, 31, 34, 203, 204, 218, 219
 - head linear, 191
 - label, 30, 150–152
 - loss, 185
 - models, 4, 8, 15, 29, 144, 176, 208, 210, 218–220
 - neural networks, 132
 - probability, 18, 31, 185, 203, 204, 218
 - rotation, 188
 - scores, 31, 218, 221, 222
 - value, 138, 175, 176
- Pretrained
 - AM, 202
 - data representations, 194
 - ImageNet classifiers, 207
 - ImageNet model, 38, 201
 - ML model, 206
 - model, 135, 203, 233
 - model parameters, 201, 210
 - representation network, 194
 - surrogate models, 38
 - voice models, 204
- Pretraining
 - datasets, 189
 - generalization ability, 191
 - methods, 198
 - phase, 191
 - setup, 197
- Primal constraints, 164

- Projected gradient descent (PGD), 195
 - adversarial training, 122
 - algorithm, 120, 234
 - attack, 19–21, 23, 73, 122, 128, 187, 195
 - attack generation method, 186
 - inner iteration, 122
 - iteration, 122
 - steps, 120, 121
 - updates, 121
- Pruned models, 227, 233, 235, 236, 238

Q

- Quadratic programming (QP), 161–164
 - problem, 161, 163, 164
- Queried data inputs, 206

R

- Random
 - noise inputs, 149
 - perturbations, 135, 238
- Random forest (RF), 157, 170
- Random number generator (RNG), 230
- Randomized
 - models, 127
 - models robustness, 127
 - smoothing, 132, 135, 137
 - smoothing technique, 135
- Rationale, 51, 150, 152, 193, 194, 202, 247
- Receiver operating characteristic (ROC), 154, 238
- Recurrent neural network (RNN), 10, 63, 66
- ReLU, 81, 89, 140
 - activations, 82
 - functions, 81, 89
 - layer, 82, 101
 - networks, 83, 89
 - neurons, 90
 - node, 91
 - relaxation in CROWN, 82
 - splits, 93
 - unit, 90
- Reprogrammed
 - input, 207
 - sample, 203
 - target data, 203, 209, 211
 - target data sample, 203

- Reprogramming
 - accuracy, 204
 - adversarial, 201
 - Clarifai, 208
 - general image models, 206
 - loss objective, 210
 - models, 201, 203, 206, 209
 - performance assessment, 212
 - voice models, 201
- Reversed layer, 247
- Robust accuracy (RA), 187, 195
- Robustness
 - adversarial, 7–10, 12, 24, 59, 68, 73, 76, 113, 115, 130, 132, 170, 183, 184, 186, 188, 189, 250
 - assessment, 232
 - bound certified, 134
 - certified, 10, 134, 135, 137, 140, 165
 - challenge, 189
 - difference, 197
 - enhancement, 191
 - evaluation, 164, 174–176, 241
 - evaluation metrics, 195
 - guarantees, 132, 137
 - improvements, 115, 187
 - measurement, 73
 - neural networks, 68, 69, 115
 - property, 73
 - regularization, 186
 - score, 170
 - transferability, 186, 194, 197
 - verification, 73, 74, 76, 85, 99, 157, 161, 162, 167, 173, 175, 176
 - algorithm, 97
 - approach, 95
 - methods, 74, 97
- Root mean squared error (RMSE), 210
- S**
- Safety verification problem, 76
- Semantic
 - adversarial attacks, 95
 - adversarial examples, 95–97
 - attack, 98, 99, 101, 107
 - attack threat, 99, 105
 - perturbation, 95, 97–100, 102, 105
 - preserved perturbations, 77
 - verification, 102
- Semantic perturbation layer (SP-layer), 97, 99
- Semisupervised machine learning, 6
- Source
 - labels, 203, 204
 - loss, 211
- Sparse
 - autoencoder, 247
 - perturbation mask, 152
 - regression models robustness, 60
- Standard
 - datasets MNIST, 41
 - robust training, 184
 - training, 121
- Standard accuracy (SA), 8, 187, 189, 195
- Standard linear finetuning (SLF), 191, 195–197
- State-of-the-art (SOTA)
 - performance, 207
 - results, 207
- Stochastic gradient descent update, 129
- Strawberry training set, 205, 206
- Superior robustness, 10
- Supervised
 - adversarial example, 243
 - attack, 245
 - attack rationale, 244
 - finetuning, 189, 190
 - machine learning, 6
 - models, 242
 - prediction, 195
- Support vector machine (SVM), 157
- T**
- t-distributed stochastic neighbor embedding (tSNE), 205
- Target
 - attack class label, 31
 - class, 17, 18, 22, 49, 144, 150, 152, 153, 206, 230
 - data inputs, 202
 - data reprogrammed, 203, 209, 211
 - dataset, 191
 - label, 52, 53, 93, 138, 150, 152–154, 203, 204, 210

- model, 15, 29–31, 38, 39, 41, 51, 52
- objects, 49
- Targeted
 - attack, 16–18, 22, 24, 31, 35, 107
 - keyword attack, 65
 - poisoning attack, 52
- Temporal dependency (TD), 146–148
 - distance, 147
 - loss, 147
- Threat model, 23, 30, 59, 73, 95, 97, 98, 102, 107, 228
- Trainable
 - additive input transformation, 203
 - input transformation function, 202
 - objective function, 168
 - parameters, 203
 - reprogram layer, 202
 - universal input perturbation, 201
- Trained
 - DNN, 149
 - model, 97, 155, 201, 208
- Transfer attack, 30
- Transferability, 38, 236
 - adversarial examples, 38, 235
 - from pretraining, 197
 - robustness, 186, 194, 197
- Transferable
 - adversarial examples, 38
 - attacks, 38
- Tree ensemble, 176, 179
 - certified robust, 181
 - models, 176
 - robustness, 179
 - verification problem, 177
- Trigger pattern, 52, 53, 56, 151
- Trojan
 - attack, 52, 149, 153
 - attack target label, 152
 - perturbation, 150
 - trigger, 150, 152
- Trojan network (TrojanNet), 149, 150, 153–155
 - detection, 149, 150, 154
- TrojanNet detector (TND), 149, 153
- True
 - gradient, 31
 - label, 185

U

- Unconstrained
 - form, 21
 - optimization formulations, 23
 - optimization problem, 20
 - problems, 21
- Unified attack formulation, 243
- Universal
 - adversarial perturbation, 22
 - attack performance, 23
 - input perturbation trainable, 201
 - perturbation, 22, 150, 152
 - perturbation generation, 152
 - trainable additive input, 207
- Unlabeled
 - data, 188, 194
 - data augmentation, 188
 - dataset, 194
 - source dataset, 190
- Unstable
 - neurons, 81, 82
 - ReLU neurons, 93
- Unsupervised
 - attack, 244
 - formulation, 244, 245
 - function, 244
 - contrastive loss, 194
 - example adversarial, 244
 - machine learning, 6
 - performance models, 248
 - tasks, 246
- Unsupervised adversarial example (UAE), 241–243, 245, 248
- Untampered models, 53
- Untargeted attack, 16–18, 21, 31, 35, 37, 41, 151, 244
- Upsampling layer, 247

V

- Vanilla adversarial reprogramming, 207
- Variational autoencoder (VAE), 221
- Verification
 - algorithms, 87, 93, 139, 159, 161
 - framework, 93
 - methods, 77, 85, 95, 114, 137, 138, 140, 180

- neural network, 76, 77
 - performance, 102
 - problem for neural network, 99
 - process, 90
 - results for datasets, 103
 - robustness, 73, 74, 76, 85, 99, 157, 161, 162, 167, 173, 175, 176
 - semantic, 102
 - techniques, 137
 - tools, 89, 95
 - Victim
 - classifier, 17
 - model, 15, 16, 22
 - Virtual adversary, 3
 - Voice to series (V2S), 202, 203
 - loss, 204
 - model architectures, 205
 - reprogramming, 204, 205
 - training, 212
- W**
- Watermark, 228–230
 - bit, 230, 231
 - carrier, 230
 - embedding, 229, 230
 - extraction, 231, 232
 - information, 229
 - Watermarking
 - bit, 231
 - digital, 227
 - technique, 228, 232
 - Weight perturbations, 67–69
 - White-box attack, 29
 - Word error rate (WER), 147, 148
 - Wrong prediction, 37, 51
- Z**
- Zeroth-order optimization (ZOO), 30, 31
 - attack, 41