Contents

1 Introduction **1**

1.1 3D Shape Completion under Weak Supervision . . . 3

1.2 Robust and Uncertainty-Aware Deep Learning . . . 4

1.2.1 Understanding Adversarial Examples and Training . . . 4

1.2.2 Improving Robustness Against Weight Perturbations . . . 6

1.2.3 Improving Adversarial Robustness and Uncertainty Estimation . . . 7

1.3 Outline . . . 8

1.4 Publications . . . 10

2 Related Work **11**

2.1 3D Shape Completion . . . 11

2.1.1 Shape Completion . . . 12

2.1.2 Shape Models . . . 12

2.1.3 Shape Representations . . . 13

2.2 Robustness and Uncertainty for Secure and Trustworthy Deep Learning . . . 13

2.3 Adversarial Robustness . . . 15

2.3.1 Adversarial Attacks . . . 15

2.3.2 Defenses Against Adversarial Attacks . . . 17

2.4 Robustness for Neural Network Accelerators . . . 20

2.4.1 Bit Error Robustness for Accelerators . . . 20

2.4.2 Quantization Robustness . . . 21

2.4.3 Weight Robustness and Fault Tolerance . . . 21

2.5 Uncertainty Estimation . . . 22

2.5.1 Out-of-Distribution Calibration and Detection . . . 22

2.5.2 Conformal Prediction . . . 23

I Deep Learning for 3D Shape Completion **25**

3 Learning 3D Shape Completion under Weak Supervision **27**

3.1 Introduction . . . 27

3.2 3D Shape Completion with Amortized Maximum Likelihood . . . 30

3.2.1 Problem Formulation . . . 30

3.2.2 Shape Prior . . . 31

3.2.3 Shape Inference . . . 32

3.2.4 Practical Considerations . . . 34

3.3 Experiments . . . 35

3.3.1 Data . . . 35

3.3.2 Evaluation . . . 37

3.3.3 Architectures and Training . . . 37

3.3.4 Baselines . . . 39

3.3.5 Experimental Evaluation . . . 40

3.4 Conclusion . . . 46

II Understanding Adversarial Examples and Training 47

4 Disentangling Adversarial Robustness and Generalization 49
4.1 Introduction 50
4.2 Adversarial Robustness and Generalization 51
4.2.1 Experimental Setup 51
4.2.2 Adversarial Examples Leave the Manifold 54
4.2.3 On-Manifold Adversarial Examples 56
4.2.4 On-Manifold Robustness is Essentially Generalization 57
4.2.5 Regular Robustness is Independent of Generalization 60
4.2.6 Discussion 61
4.3 Conclusion 63

5 Relating Adversarially Robust Generalization to Flat Minima 65
5.1 Introduction 66
5.2 Robust Generalization and Flat Minima 67
5.2.1 Adversarial Training and Robust Overfitting 68
5.2.2 Intuition and Visualizing Flatness 69
5.2.3 Average- and Worst-Case Flatness Measures 70
5.2.4 Discussion 71
5.3 Experiments 72
5.3.1 Understanding Robust Overfitting 74
5.3.2 Robust Generalization and Flatness 75
5.4 Conclusion 79

III Improving Weight Robustness 81

6 Random and Adversarial Bit Error Robustness 83
6.1 Introduction 84
6.2 Bit Errors in Quantized Weights 87
6.2.1 Low-Voltage Induced Random Bit Errors 87
6.2.2 Adversarial Bit Errors 89
6.3 Robustness Against Bit Errors 91
6.3.1 Robust Fixed-Point Quantization 91
6.3.2 Training with Weight Clipping as Regularization 93
6.3.3 Random Bit Error Training (RandBET) 95
6.3.4 Adversarial Bit Error Training (AdvBET) 96
6.4 Experiments 97
6.4.1 Setup and Baselines 97
6.4.2 Robust Test Error and Generalization Bound 98
6.4.3 Batch Normalization is not Robust 100
6.4.4 Quantization Choice Impacts Robustness 100
6.4.5 Weight Clipping Improves Robustness 101
6.4.6 RandBET Yields Generalizable Robustness 103
6.4.7 Robustness to Bit Errors in Inputs and Activations 106
6.4.8 Robustness Against Adversarial Bit Errors 108
6.5 Conclusion 111

IV Improving Adversarial Robustness and Uncertainty Estimation 113

7 Confidence-Calibrated Adversarial Training 115
7.1 Introduction . . . 116
7.2 Confidence Calibration of Adversarial Examples . . . 117
7.2.1 Problems of Adversarial Training . . . 117
7.2.2 Confidence-Calibrated Adversarial Training . . . 119
7.3 Detection and Robustness Evaluation with Adaptive Attack . . . 124
7.3.1 Adaptive Attack . . . 124
7.3.2 Detection Evaluation . . . 124
7.3.3 Robustness Evaluation . . . 125
7.3.4 Per-Example Worst-Case Evaluation . . . 126
7.4 Experiments . . . 126
7.4.1 Attacks . . . 126
7.4.2 Training and Baselines . . . 128
7.4.3 Ablation Study . . . 129
7.4.4 Main Results . . . 129
7.4.5 Analysis . . . 133
7.5 Conclusion . . . 134
7.5.1 Discussion of Recent Results . . . 134

8 Learning Optimal Conformal Classifiers 135
8.1 Introduction . . . 136
8.2 Differentiable Conformal Predictors . . . 137
8.2.1 Conformal Predictors . . . 138
8.2.2 Differentiable Prediction and Calibration Steps . . . 139
8.3 Conformal Training: Learning Conformal Prediction . . . 140
8.3.1 Conformal Training by Optimizing Inefficiency . . . 140
8.3.2 Conformal Training with Classification Loss . . . 141
8.3.3 Conformal Training with General and Application-Specific Losses . . . 142
8.3.4 Conformal Training as Generalization of Coverage Training . . . 143
8.4 Experiments . . . 144
8.4.1 Experimental Setup . . . 144
8.4.2 Reducing Inefficiency with Conformal Training . . . 145
8.4.3 Conformal Training for Applications: Case Studies . . . 147
8.5 Conclusion . . . 149

9 Conclusion and Future Work 151
9.1 Key Insights and Conclusions . . . 151
9.2 Future Directions . . . 153

List of Algorithms 157

List of Figures 159

List of Tables 163

Bibliography 165

A Disentangling Adversarial Robustness and Generalization 227

A.1 On-Manifold Adversarial Examples . . . 227
A.2 L_2 and Transfer Attacks . . . 227
A.3 Influence of Network Architecture . . . 230
A.4 Baselines and Adversarial Training Variants . . . 231

B Relating Adversarially Robust Generalization to Flat Minima 233
B.1 Visualization Details and Discussion . . . 233
B.2 Ablation for Flatness Measures . . . 233
B.3 Scaling Networks and Scale-Invariance . . . 235
B.4 Methods . . . 236
B.4.1 Training Curves . . . 239
B.4.2 Flatness . . . 240
B.5 Results in Tabular Form . . . 241

C Random and Adversarial Bit Error Robustness 245
C.1 Complete Profiled Bit Error Statistics and Evaluation . . . 245
C.2 Quantization and Bit Manipulation in PyTorch . . . 246
C.3 Network Architectures and Expected Bit Errors . . . 247
C.4 Additional Experiments . . . 247
C.4.1 Architecture Comparison . . . 247
C.4.2 Robust Quantization . . . 248
C.4.3 (Global) Weight Clipping (Clipping) . . . 249
C.4.4 Random Bit Error Training (RandBET) . . . 250
C.4.5 Per-Layer Clipping and RandBET . . . 252
C.4.6 Profiled Bit Errors . . . 252
C.4.7 Summary Results . . . 253
C.4.8 Bit Errors in Activations and Inputs . . . 255
C.4.9 Adversarial Bit Error Attack . . . 256

D Confidence-Calibrated Adversarial Training 259
D.1 PGD with Momentum and Backtracking . . . 259
D.2 Attack Initialization . . . 260
D.3 ROC AUC and Computing Confidence Threshold . . . 260
D.3.1 Ablation Study . . . 261
D.4 All Results . . . 263

E Learning Optimal Conformal Classifiers 267
E.1 Experimental Setup . . . 267
E.2 Importance of Random Trials . . . 270
E.3 Coverage/Conformal Training Ablation on Fashion-MNIST . . . 271
E.4 Impact of Hyperparameters . . . 271
E.5 All Inefficiency Results . . . 272
E.6 Shaping Class-Conditional Inefficiency on Other Datasets . . . 273
E.7 Manipulating Coverage Confusion on Other Datasets . . . 274
E.8 Miscoverage Results on Additional Datasets . . . 274
E.9 Additional Results on Binary Datasets . . . 275
E.10 Effect of Conformal Training on Class-Conditional Inefficiency and Coverage Confusion . . . 276