Contents

List of Figures *page* ix
List of Tables xi

1 Introduction 1
1.1 Background 1
1.2 From Cloud Learning to Edge Learning 2
1.2.1 From Cloud Computing to Edge Computing 2
1.2.2 From Distributed Machine Learning to Edge Learning 3
1.3 Edge Learning and Edge Intelligence 4
1.4 Challenges of Edge Learning 5
1.4.1 Hard to Train Due to Constrained and Heterogeneous Edge Resources 6
1.4.2 Hard to Protect Due to Vulnerable Edge Devices 7
1.4.3 Hard to Manage Due to Complex Edge Environment 8
1.4.4 Hard to Collaborate Due to Lack of Participant 8
1.5 The Scope and Organization of This Book 9

2 Preliminary 11
2.1 Background of Edge Computing 11
2.1.1 Edge Computing Paradigms 11
2.2 Deep Learning Models and Collaborative Training Approaches 14
2.2.1 Deep Learning Models 14
2.2.2 Collaborative Training Approaches 15
2.3 Basic Machine Learning Algorithms 16
2.3.1 Learning Problem Statement 16
2.3.2 Basic Machine Learning Algorithms 17
2.4 Learning Architectures: Parameter Server and Decentralized Learning 20
2.5 Synchronization Modes 22
2.5.1 Bulk Synchronous Parallel (BSP) 23
2.5.2 Asynchronous Parallel (ASP) 23
2.5.3 Stale Synchronous Parallel (SSP) 23

Cambridge University Press
978-1-108-83237-3 — Edge Learning for Distributed Big Data Analytics
Song Guo , Zhihao Qu
Table of Contents
More Information

3 Fundamental Theory and Algorithms of Edge Learning 24
3.1 Distributed Machine Learning and the Convergence Theory 24
3.2 Advanced Training Algorithm and Corresponding Theory 27
3.2.1 Regularization and Loss Function 27
3.2.2 Direction Based Optimization 29
3.2.3 Algorithms Based on Hyper-Parameters 32
3.2.4 Co-designed Algorithms 34
3.2.5 Optimization Algorithms for DNN 37
3.3 Theoretical Framework for Flexible Synchronization in Edge Learning 38

4 Communication-Efficient Edge Learning 42
4.1 Introduction to Communication-Efficient Edge Learning 42
4.2 Communication Data Compression in Edge Learning 43
4.2.1 Quantization 43
4.2.2 Sparsification 46
4.2.3 Low Rank 48
4.2.4 Error Compensation Techniques for Communication Compression 48
4.2.5 Communication Compression in Decentralized Training 50
4.3 Lazy Synchronization 51
4.3.1 Large Batch Size 51
4.3.2 Periodic Averaging 52
4.3.3 Fine-Grained Aggregation 52
4.3.4 A Communication-Efficient Edge Learning Framework with Quantized and Period Averaging 53
4.4 Overlap Synchronization Parallel with Quantization 57
4.4.1 Algorithm Description 58
4.4.2 Theoretical Results 60
4.5 Wireless Network Optimization for Edge Learning 62
4.5.1 Scheduling Policy for Communication-Efficient Edge Learning in Wireless Environments 63
4.5.2 MIMO and Over-the-Air Computation for Fast Aggregation in Edge Learning 66
4.6 Conclusion and Future Directions 70
4.6.1 Two-Pass Compression Method for Edge Learning 71
4.6.2 Gradient Compression Robust to Byzantine Workers 71
4.6.3 Communication Compression for Two-Order Optimization Algorithm 72

5 Computation Acceleration 73
5.1 Introduction to Computation Acceleration 73
5.2 Model Compression and Hardware Acceleration 74
5.2.1 model compression 74
5.2.2 Hardware Acceleration 79

CAMBRIDGE

Cambridge University Press
978-1-108-83237-3 — Edge Learning for Distributed Big Data Analytics
Song Guo , Zhihao Qu
Table of Contents
More Information

5.3 Straggler Tolerance 82
5.3.1 Framework of Gradient Coding 83
5.3.2 Construction Encoding and Decoding Matrix 84
5.3.3 Construct **B** in the General Case 87
5.3.4 Recent Methods of Gradient Coding 90
5.4 Improving the Inference Performance in the Edge Environment 91
5.4.1 Key Performance Indicators in Inference 91
5.4.2 Enabling Technologies for Inference 92
5.5 Conclusion and Future Directions 95
5.5.1 Jointly Optimize Learning Algorithm and Hardware Implementation in Edge Environments 95
5.5.2 Green and Sustainable Model Training among Heterogeneous Hardware Platforms 96
5.5.3 Approximate Gradient Coding to Deal with Stragglers 97

6 Efficient Training with Heterogeneous Data Distribution 98
6.1 Introduction to Federated Learning 98
6.2 Training with Non-IID Data 101
6.2.1 What Does Non-IID Mean? 102
6.2.2 Enabling Technologies for Training Non-IID Data 102
6.3 Conclusion and Future Directions 107
6.3.1 Tackle the Non-IID Data via Learning-based Data Selection 108
6.3.2 Adaptive Parameter Setting for Non-IID Data 109
6.3.3 Straggler-Tolerant Federated Learning Algorithms 110

7 Security and Privacy Issues in Edge Learning Systems 112
7.1 Security Guarantee 112
7.1.1 Data-Oriented Attacks 113
7.1.2 Defense Technologies for Data-Oriented Attacks 116
7.1.3 Model-Oriented Attacks 119
7.1.4 Defense Technologies for Model-Oriented Attacks 120
7.2 Privacy Protection 121
7.2.1 Introduction to Privacy Attacks in Edge Learning 121
7.2.2 Enabling Technologies for Private Edge Learning 122
7.3 Conclusion and Future Directions 128
7.3.1 Multi-level Privacy-Protection for Efficient Edge Learning 128
7.3.2 Hierarchical Outlier Detection for Security Guarantee 128
7.3.3 Attack Detection in Communication-Compressed Training 130
7.3.4 Computation Offloading for Encrypted Data Training 130

8 Edge Learning Architecture Design for System Scalability 131
8.1 Introduction to the Learning Architecture 131
8.1.1 Parallelism Schemes: Data Parallelism and Model Parallelism 131

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-83237-3 — Edge Learning for Distributed Big Data Analytics
Song Guo , Zhihao Qu
Table of Contents
More Information

8.1.2 Large-Scale Model Training Architecture 137
8.2 Edge Learning Frameworks over the Hierarchical Architecture 139
8.2.1 Introduction to the Hierarchical Architecture 140
8.2.2 Community-Based Synchronization Parallel over the Hierarchical Architecture 143
8.2.3 Convergence Rate of Community-Based Synchronization Parallel 145
8.3 Extension of Community-Based Synchronization Parallel 150
8.3.1 A Hybrid Synchronization Mechanism over the Hierarchical Architecture 150
8.3.2 Abstract of Community and Communication-Aware Parameter Servers 151
8.3.3 Convergence Result of Hybrid Community-Based Synchronization Parallel 152
8.4 Conclusion and Future Directions 157

9 Incentive Mechanisms in Edge Learning Systems 159
9.1 Fundamental Theory of Incentive Mechanisms 159
9.2 Related Works 161
9.2.1 Incentive Mechanisms 161
9.2.2 Incentive Mechanisms for Edge Learning 161
9.3 A Learning-Based Incentive Mechanism for Edge Learning 162
9.3.1 Problem Description 164
9.3.2 System Model 165
9.3.3 Equilibrium Analysis 166
9.3.4 A Deep Reinforcement Learning-Based Incentive Mechanism 168
9.4 Conclusion and Future Directions 169

10 Edge Learning Applications 171
10.1 APIs, Libraries, and Platforms for Edge Learning 171
10.1.1 General Programming Frameworks for Machine Learning 171
10.2 Application Scenarios 175
10.2.1 Smart Transportation 175
10.2.2 Smart Healthcare 179
10.2.3 Intelligent Blockchain + Edge AI 180
10.2.4 Intelligent Financial Risk Control 182
10.2.5 Edge AI + IoT 184
10.2.6 Virtual Reality 186
10.3 The Dr. Body System for Posture Detection and Rehabilitation Tracking 188

Bibliography 190
Index 215

© in this web service Cambridge University Press www.cambridge.org