Edge Learning for Distributed Big Data Analytics

Discover this multi disciplinary and insightful work, which integrates machine learning, edge computing, and big data. It presents the basics of training machine learning models, key challenges and issues, as well as comprehensive techniques including edge learning algorithms and system design issues. The volume describes architectures, frameworks, and key technologies for learning performance, security, and privacy, as well as incentive issues in training/inference at the network edge. It is intended to stimulate fruitful discussions, inspire further research ideas, and inform readers from academia and those having an industry background. It is an essential read for experienced researchers and developers, as well as for those who are just entering the field.

Song Guo is a Full Professor in the Department of Computing at the Hong Kong Polytechnic University. He is an IEEE Fellow and the editor-in-chief of the *IEEE Open Journal of the Computer Society*. He was a member of the IEEE ComSoc Board of Governors and a distinguished lecturer of the IEEE Communications Society.

Zhihao Qu is an assistant researcher in the School of Computer and Information at Hohai University and in the Department of Computing at the Hong Kong Polytechnic University.

The authors of this book would like to acknowledge the following contributors: Jie Zhang for contributing materials to Chapter 6 (Efficient Training with Heterogeneous Data Distribution), Qihua Zhou for contributing materials to Chapter 8 (Edge Learning Architecture Design for System Scalability), Haozhao Wang for contributing materials to Chapter 5 (Computation Acceleration), Yufeng Zhan for contributing materials to Chapter 9 (Incentive Mechanisms in Edge Learning Systems), and Chenxi Chen for contributing materials to Chapter 7 (Security and Privacy Issues in Edge Learning Systems).

Edge Learning for Distributed Big Data Analytics

Theory, Algorithms, and System Design

SONG GUO The Hong Kong Polytechnic University

ZHIHAO QU Hohai University and The Hong Kong Polytechnic University



© in this web service Cambridge University Press



University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314-321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi - 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org Information on this title: www.cambridge.org/9781108832373 DOI: 10.1017/9781108955959

© Cambridge University Press 2022

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2022

Printed in the United Kingdom by TJ Books Ltd, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

ISBN 978-1-108-83237-3 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

1

2

Cambridge University Press 978-1-108-83237-3 — Edge Learning for Distributed Big Data Analytics Song Guo , Zhihao Qu Frontmatter <u>More Information</u>

Contents

List	t of Figures	<i>page</i> ix
List	t of Tables	xi
Intro	oduction	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 Lea	arning 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 Heterogen	ieous
	Edge Resources	6
	1.4.2 Hard to Protect Due to Vulnerable Edge Devices	
	1.4.3 Hard to Manage Due to Complex Edge Environment	ment 8
	1.4.4 Hard to Collaborate Due to Lack of Participant	8
1.5	The Scope and Organization of This Book	9
Preli	liminary	11
2.1	Background of Edge Computing	11
	2.1.1 Edge Computing Paradigms	11
2.2	Deep Learning Models and Collaborative Training Appr	roaches 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 Decentral	lized 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

vi	Cont	ents	
3	Func	lamental 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	Com	munication-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	50
		Framework with Quantized and Period Averaging	53
	4.4	Overlap Synchronization Parallel with Quantization	57
		4.4.1 Algorithm Description	58
	4.5	4.4.2 Theoretical Results	60
	4.5	Wireless Network Optimization for Edge Learning	62
		4.5.1 Scheduling Policy for Communication-Efficient Edge	(2
		4.5.2 MIMO and Quar the Air Computation for Fast	03
		4.5.2 MINO and Over-the-Air Computation for Fast	66
	16	Conclusion and Future Directions	00 70
	4.0	4.6.1 Two Pass Compression Method for Edge Learning	70
		4.6.1 Two-rass Compression Method for Edge Learning	71
		4.6.2 Communication Compression for Two Order	/1
		Optimization Algorithm	72
5	Com	nutation Acceleration	73
~	5 1	Introduction to Computation Acceleration	73
	5.1	Model Compression and Hardware Acceleration	74
	5.2	5.2.1 model compression	74
		5.2.2 Hardware Acceleration	, t 79
			1)

		Contents	s vii
	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	Effic	sient 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	Secu	urity 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	e Learning Architecture Design for System Scalability	131
	8.1	Introduction to the Learning Architecture	131
		8.1.1 Parallelism Schemes: Data Parallelism and Model Parallelis	m 131

viii	Cont	ents	
		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	1.50
		Synchronization Parallel	152
	8.4	Conclusion and Future Directions	157
9	Ince	ntive 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	e 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
	Bibl	iography	190
	Inde	x	215

Figures

1.1	Cloud-Edge environment	page 4
1.2	Summary of the main challenges in edge learning.	5
1.3	Communication, computation, and data challenge in edge learning	6
1.4	Privacy and Security challenge	7
1.5	Challenges in edge device collaboration.	8
1.6	The scope and organization of this book.	9
2.1	The architecture of a cloudlet.	12
2.2	Illustration of a micro data center.	13
2.3	Illustration of a fog computing framework.	13
2.4	Illustration of mobile edge computing.	14
2.5	Example of different architectures.	21
2.6	Illustration of different synchronization modes.	22
4.1	Gradient quantization method.	44
4.2	The basic idea of sparsification method.	47
4.3	Illustration of original distributed SGD, OSP, and QOSP.	59
4.4	Communication-efficient scheduling policy in wireless environments	64
4.5	Illustration of over-the-air computation for fast aggregation in edge	
	learning.	68
5.1	Illustration of low rank factorization	75
5.2	Network pruning.	76
5.3	Illustration of gradient coding methods for tolerating stragglers.	83
5.4	Illustration of the high-level deep learning workflow	91
5.5	The process of pruning	93
5.6	Example scenario of computation offloading	95
6.1	An illustration of federated learning.	100
6.2	Illustration of the weight divergence for different data distribution	103
6.3	Illustration of the data-sharing strategy.	104
6.4	Accompanying model for training data selection.	108
6.5	Adaptive parameter setting for non-IID data in federated learning.	110
7.1	One example of data poisoning attacks	115
7.2	An example of model attacks	119
7.3	An illustration of homomorphic encryption	126
8.1	Data parallelism.	132
8.2	Model parallelism.	134

x List of Figures	
-------------------	--

8.3	Illustration of hybrid parallelism.	135
8.4	The dataflow of training MNIST via Keras.	139
8.5	Example of community-based synchronization parallel	141
9.1	An illustration of mechanism design.	160
9.2	Test accuracy with varying the size of training data on MNIST dataset.	163
10.1	Illustration of smart transportation platform.	178
10.2	An example of edge-based healthcare application.	180
10.3	An example of edge-based Blockchain.	181
10.4	An example of intelligent risk control system.	184
10.5	An example of edge-based IoT framework.	185
10.6	Intelligent video analysis framework in edge environment.	187

Tables

4.1	1 Optimization objectives of scheduling policies in edge learning with	
	constrained resources	page 67
6.1	Comparison of distributed learning and federated learning.	99
7.1	Summary of data-oriented and model-oriented defense methods	114
7.2	Summary of privacy-preserving methods	129