

Xiaofei Wang · Yiwen Han
Victor C. M. Leung · Dusit Niyato
Xueqiang Yan · Xu Chen

Edge AI

Convergence of Edge Computing and
Artificial Intelligence



Springer

Edge AI

Xiaofei Wang • Yiwen Han • Victor C. M. Leung •
Dusit Niyato • Xueqiang Yan • Xu Chen

Edge AI

Convergence of Edge Computing
and Artificial Intelligence



Springer

Xiaofei Wang
College of Intelligence and Computing
Tianjin University
Tianjin, Tianjin, China

Yiwen Han
College of Intelligence and Computing
Tianjin University
Tianjin, Tianjin, China

Victor C. M. Leung
College of Computer Science and Software
Engineering
Shenzhen University
Shenzhen, Guangdong, China

Dusit Niyato
School of Computer and Engineering
Nanyang Technological University
Singapore, Singapore

Xueqiang Yan
2012 Lab
Huawei Technologies (China)
Shenzhen, China

Xu Chen
School of Data and Computer Science
Sun Yat-sen University
Guangzhou, Guangdong, China

ISBN 978-981-15-6185-6 ISBN 978-981-15-6186-3 (eBook)

<https://doi.org/10.1007/978-981-15-6186-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. 2020

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

At present, we are living in an era of rapid development. Artificial intelligence (AI), as a technology leading the trend of this era and subverting people's traditional lifestyle, is deeply integrated into production and life around the world. With its rapid rise in the fields of smart factories, smart cities, smart homes, and smart Internet of Things, it uses technology to achieve the interaction between humans and machines. At present, AI has replaced human roles in many key areas of high intensity, difficulty, and danger and even has surpassed human capabilities in some areas. Therefore, AI has promoted the liberation of human freedom to a certain extent by replacing labor and management of human beings.

However, the realization of AI requires large-scale data as support. It is based on the training and learning of a large number of sample data that AI can perform nearly or even surpass human performance. The current data in the network is showing an exponential growth, which has created opportunities for the rise and development of AI. However, the rapid increase in the size of network data is a great challenge of current network architecture. In order to alleviate the pressure on the network caused by the explosive growth of data scale, edge computing technology came into being. Edge computing can reduce the pressure on the network and reduce the delay in request response by setting distributed edge nodes at the edge of the network.

The explosive growth of data not only provides a prerequisite for the development of AI but also creates opportunities for the rise of edge computing. However, AI and edge computing, as two popular emerging technologies, are inextricably linked to each other. On the one hand, the characteristics of edge computing that can reduce latency and traffic load can provide basic guarantees for AI. On the other hand, the learning and decision-making capabilities of AI are supporting the efficient and stable operation of edge computing. The two technologies are not only mutually supporting but also merging with each other, and they are inseparable. Deep learning, as the most representative technology of AI in combination with edge computing, has made remarkable progress in many fields through cooperation with edge computing. In this background, the purpose of this monograph is to explore the

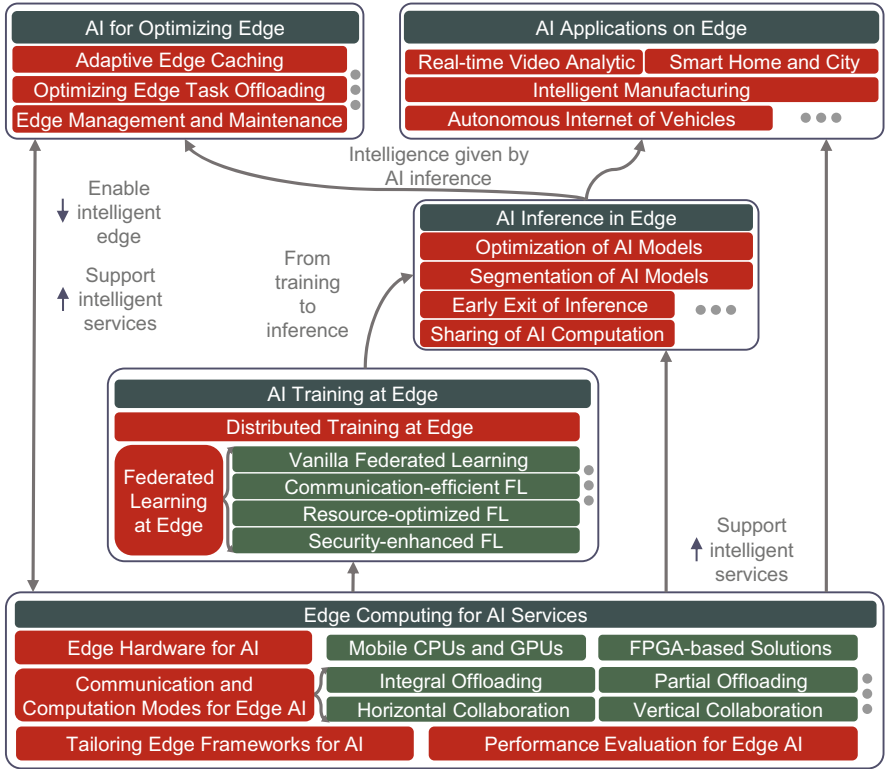


Fig. 1 Conceptual relationships of edge intelligence and intelligent edge

relevant achievements around the relationship between edge computing and artificial intelligence.

This monograph introduces and discusses the advanced technology of edge AI in terms of fundamentals, concepts, frameworks, application cases, optimization method, and future directions, so as to provide students, researchers, and practitioners in related fields with a comprehensive reference. In detail, this book is organized as follows (as abstracted in Fig. 1): In Chap. 1, we have introduced the generation, development, trend, and industrial status of edge computing and given the brief of intelligent edge and edge intelligence. Next, we provide some fundamentals related to edge computing and AI in Chaps. 2 and 3, respectively. The following sections introduce the five enabling technologies, i.e., AI applications on edge (Chap. 4), AI inference in edge (Chap. 5), AI training at edge (Chap. 6), edge computing for AI (Chap. 7), and AI for optimizing edge (Chap. 8). Finally, we present lessons learned and discuss open challenges in Chap. 9 and conclude this book in Chap. 10. Therefore, the intended audience includes scientific researchers and industry professionals engaged in the field of edge computing and artificial

intelligence. Hopefully, this monograph can fill in the gaps in the architecture of edge AI and further expand the existing knowledge system in this field.

Tianjin, China
 Tianjin, China
 Shenzhen, China
 Singapore, Singapore
 Shenzhen, China
 Guangzhou, Guangdong, China

Xiaofei Wang
 Yiwen Han
 Victor C. M. Leung
 Dusit Niyato
 Xueqiang Yan
 Xu Chen

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2019YFB2101901 and No.2018YFC0809803), National Science Foundation of China (No. 61702364, No. 61972432, and No. U1711265), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (No.2017ZT07X355), Chinese National Engineering Laboratory for Big Data System Computing Technology. It was also supported in part by Singapore NRF National Satellite of Excellence, Design Science and Technology for Secure Critical Infrastructure NSoE DeST-SCI2019-0007, A*STAR-NTU-SUTD Joint Research Grant Call on Artificial Intelligence for the Future of Manufacturing RGANS1906, WASP/NTU M4082187 (4080), Singapore MOE Tier 1 2017-T1-002-007 RG122/17, MOE Tier 2 MOE2014-T2-2-015 ARC4/15, Singapore NRF2015-NRF-ISF001-2277, and Singapore EMA Energy Resilience NRF2017EWT-EP003-041.

Contents

Part I Introduction and Fundamentals

1	Introduction	3
1.1	A Brief Introduction to Edge Computing	3
1.2	Trends in Edge Computing	6
1.3	Industrial Applications of Edge Computing	7
1.4	Intelligent Edge and Edge Intelligence	8
	References	12
2	Fundamentals of Edge Computing	15
2.1	Paradigms of Edge Computing	15
2.1.1	Cloudlet and Micro Data Centers	16
2.1.2	Fog Computing	17
2.1.3	Mobile and Multi-Access Edge Computing (MEC)	17
2.1.4	Definition of Edge Computing Terminologies	18
2.1.5	Collaborative End–Edge–Cloud Computing	18
2.2	Hardware for Edge Computing	19
2.2.1	AI Hardware for Edge Computing	19
2.2.2	Integrated Commodities Potentially for Edge Nodes	21
2.3	Edge Computing Frameworks	22
2.4	Virtualizing the Edge	25
2.4.1	Virtualization Techniques	27
2.4.2	Network Virtualization	28
2.4.3	Network Slicing	28
2.5	Value Scenarios for Edge Computing	29
2.5.1	Smart Parks	29
2.5.2	Video Surveillance	30
2.5.3	Industrial Internet of Things	30
	References	30
3	Fundamentals of Artificial Intelligence	33
3.1	Artificial Intelligence and Deep Learning	33

3.2	Neural Networks in Deep Learning	35
3.2.1	Fully Connected Neural Network (FCNN)	35
3.2.2	Auto-Encoder (AE)	36
3.2.3	Convolutional Neural Network (CNN)	36
3.2.4	Generative Adversarial Network (GAN)	37
3.2.5	Recurrent Neural Network (RNN)	38
3.2.6	Transfer Learning (TL)	39
3.3	Deep Reinforcement Learning (DRL)	41
3.3.1	Reinforcement Learning (RL)	41
3.3.2	Value-Based DRL	42
3.3.3	Policy-Gradient-Based DRL	42
3.4	Distributed DL Training	42
3.4.1	Data Parallelism	43
3.4.2	Model Parallelism	44
3.5	Potential DL Libraries for Edge	44
	References	45

Part II Artificial Intelligence and Edge Computing

4	Artificial Intelligence Applications on Edge	51
4.1	Real-time Video Analytic	51
4.1.1	Machine Learning Solution	51
4.1.2	Deep Learning Solution	52
4.2	Autonomous Internet of Vehicles (IoVs)	54
4.2.1	Machine Learning Solution	54
4.2.2	Deep Learning Solution	55
4.3	Intelligent Manufacturing	56
4.3.1	Machine Learning Solution	57
4.3.2	Deep Learning Solution	58
4.4	Smart Home and City	59
4.4.1	Machine Learning Solution	60
4.4.2	Deep Learning Solution	61
	References	62
5	Artificial Intelligence Inference in Edge	65
5.1	Optimization of AI Models in Edge	65
5.1.1	General Methods for Model Optimization	66
5.1.2	Model Optimization for Edge Devices	67
5.2	Segmentation of AI Models	69
5.3	Early Exit of Inference (EEoI)	71
5.4	Sharing of AI Computation	72
	References	74
6	Artificial Intelligence Training at Edge	77
6.1	Distributed Training at Edge	77
6.2	Vanilla Federated Learning at Edge	82
6.3	Communication-Efficient FL	83

6.4	Resource-Optimized FL	85
6.5	Security-Enhanced FL	86
6.6	A Case Study for Training DRL at Edge	89
6.6.1	Multi-User Edge Computing Scenario	89
6.6.2	System Formulation	90
6.6.3	Offloading Strategy for Computing Tasks Based on DRL	92
6.6.4	Distributed Cooperative Training	93
	References	93
7	Edge Computing for Artificial Intelligence	97
7.1	Edge Hardware for AI	97
7.1.1	Mobile CPUs and GPUs	97
7.1.2	FPGA-Based Solutions	99
7.1.3	TPU-Based Solutions	100
7.2	Edge Data Analysis for Edge AI	101
7.2.1	Challenge and Needs for Edge Data Process	101
7.2.2	Combination of Big Data and Edge Data Process	102
7.2.3	Architecture for Edge Data Process	103
7.3	Communication and Computation Modes for Edge AI	103
7.3.1	Integral Offloading	103
7.3.2	Partial Offloading	104
7.3.3	Vertical Collaboration	106
7.3.4	Horizontal Collaboration	108
7.4	Tailoring Edge Frameworks for AI	110
7.5	Performance Evaluation for Edge AI	112
	References	113
8	Artificial Intelligence for Optimizing Edge	117
8.1	AI for Adaptive Edge Caching	117
8.1.1	Use Cases of DNNs	121
8.1.2	Use Cases of DRL	122
8.2	AI for Optimizing Edge Task Offloading	123
8.2.1	Use Cases of DNNs	124
8.2.2	Use Cases of DRL	125
8.3	AI for Edge Management and Maintenance	126
8.3.1	Edge Communication	126
8.3.2	Edge Security	127
8.3.3	Joint Edge Optimization	128
8.4	A Practical Case for Adaptive Edge Caching	129
8.4.1	Multi-BS Edge Caching Scenario	129
8.4.2	System Formulation	130
8.4.3	Weighted Distributed DQN Training and Cache Replacement	131
8.4.4	Conclusion for Edge Caching Case	132
	References	132

Part III Challenges and Conclusions

9	Lessons Learned and Open Challenges	137
9.1	More Promising Applications	137
9.2	General AI Model for Inference	138
9.2.1	Ambiguous Performance Metrics	138
9.2.2	Generalization of EEoI	139
9.2.3	Hybrid Model Modification	139
9.2.4	Coordination Between AI Training and Inference	140
9.3	Complete Edge Architecture for AI	140
9.3.1	Edge for Data Processing	140
9.3.2	Microservice for Edge AI Services	142
9.3.3	Incentive and Trusty Offloading Mechanism for AI	142
9.3.4	Integration with “AI for Optimizing Edge”	143
9.4	Practical Training Principles at Edge	143
9.4.1	Data Parallelism Versus Model Parallelism	144
9.4.2	Training Data Resources	144
9.4.3	Asynchronous FL at Edge	145
9.4.4	Transfer Learning-Based Training	145
9.5	Deployment and Improvement of Intelligent Edge	146
	References	147
10	Conclusions	149

Acronyms

A3C	Asynchronous advantage actor-critic
AC	Actor-critic
AE	Auto-encoder
AI	Artificial intelligence
A-LSH	Adaptive locality sensitive hashing
ALU	Arithmetic and logic unit
APU	AI processing unit
AR	Augmented reality
ASIC	Application specific integrated circuit
B/S	Browser/server
BNNS	Binarized neural networks
BS	Base station
C/S	Client/server
CDN	Content delivery network
CNN	Convolutional neural network
C-RAN	Cloud-radio access networks
CV	Computer vision
D2D	Device-to-device
DAD	Deep architecture decomposition
DAG	Directed acyclic graph
DBMS	Database management system
DDNNs	Distributed deep neural networks
DDoS	Distributed denial of service
DDPG	Deep deterministic policy gradient
DL	Deep learning
DNN	Deep neural network
Double-DQL	Double deep Q -learning
DP	Differential privacy
DQL	Deep Q -learning
DQN	Deep Q -learning network
DRAM	Dynamic RAM

DRL	Deep reinforcement learning
DSL	Domain-specific language
Dueling-DQL	Dueling deep Q -learning
DVFS	Dynamic voltage and frequency scaling
eBNNs	Embedded binarized neural networks
ECC	Edge Computing Consortium
ECSP	Edge computing service provider
EEoI	Early exit of inference
EH	Energy harvesting
ETSI	European Telecommunications Standards Institute
FAP	Fog radio access point
FCN	Fog computing node
FCNN	Fully connected neural network
FIFO	First in first out
FL	Federated learning
FTP	Fused Tile Partitioning
GAN	Generative adversarial network
GNN	Graph neural network
GPU	Graphics processing unit
IID	Independent and identically distributed
IOB	Input–output block
IoT	Internet of Things
IoVs	Internet of vehicles
KD	Knowledge distillation
k NN	k -nearest neighbor
LCA	Logic cell array
LFU	Least frequently used
LRU	Least recently used
LSTM	Long short-term memory
MAB	Multi-armed bandit
MDC	Micro data center
MDP	Markov decision process
MEC	Mobile (multi-access) edge computing
MLP	Multi-layer perceptron
NCS	Neural compute stick
NFV	Network functions virtualization
NLP	Natural language processing
NN	Neural network
NPU	Neural processing unit
P2P	Peer-to-peer
PC	Personal computer
PPO	Proximate policy optimization
QoE	Quality of experience
QoS	Quality of service
RAM	Random access memory

RAN	Radio access network
RL	Reinforcement learning
RNN	Recurrent neural network
RoI	Region-of-Interest
RPC	Remote procedure call
RRH	Remote radio head
RSU	Road-side unit
SDK	Software development kit
SDN	Software-defined network
SGD	Stochastic Gradient Descent
SINR	Signal-to-interference-plus-noise ratio
SNPE	Snapdragon neural processing engine
SNR	Signal-to-noise ratios
SRAM	Static random access memory
SVD	Singular value decomposition
TL	Transfer learning
TPU	Tensor processing unit
UE	User equipment
V2V	Vehicle-to-vehicle
VHDL	Very-high-speed integrated circuit hardware description language
VM	Virtual machine
VNF	Virtual network function
VPU	Vision processing unit
VR	Virtual reality
WLAN	Wireless local area network
ZB	Zettabytes

Part I
Introduction and Fundamentals