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

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

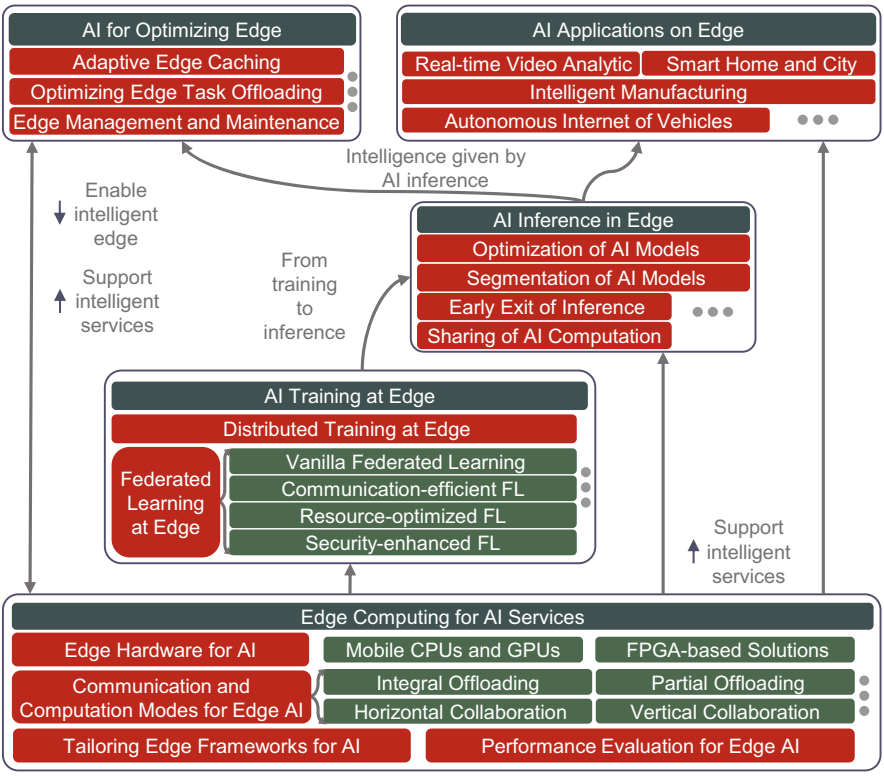

Fig. 1 Conceptual relationships of edge intelligence and intelligent edge

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 — Xiaofei Wang
Tianjin, China — Yiwen Han
Shenzhen, China — Victor C. M. Leung
Singapore, Singapore — Dusit Niyato
Shenzhen, China — Xueqiang Yan
Guangzhou, Guangdong, China — 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

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
kNN	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