Guest Editorial Special Issue on Artificial-Intelligence-Powered Edge Computing for Internet of Things

RECENT years have witnessed the proliferation of mobile computing and the Internet of Things (IoT), in which billions of mobile and IoT devices are connected to the Internet, generating zillions bytes of data at the network edge. However, it is challenging and infeasible to transfer and process zillions bytes of data using the current cloud-device architecture, due to bandwidth constraints of networks, potentially uncontrollable latency of cloud services, and privacy concerns while collecting data from IoT devices. To tackle these challenges, edge computing, an emerging computing paradigm, has received a tremendous amount of interest. By pushing data storage, computing, and controls closer to the network edge, edge computing has been widely recognized as a promising solution to meet the requirements of low latency, high scalability, and energy efficiency, as well as to mitigate the network traffic burdens. However, with the emergence of diverse IoT applications (e.g., smart city, industrial automation, and connected car), it becomes challenging for edge computing to deal with these heterogeneous IoT environments.

Motivated by the success of artificial intelligence (AI) in a wide range of fields (e.g., computer vision, speech recognition, natural language processing, chess playing (e.g., AlphaGo), and robotics), it is envisaged that AI-powered edge computing could overcome the emerging challenges and fully unleash the potential of the edge big data. The resulted new interdiscipline, edge AI or edge intelligence, is beginning to receive a tremendous amount of interest. However, integrating the AI into the edge is highly nontrivial, due to the concerns on performance, cost, and privacy. Traditional approaches for AI applications, which analyze the data bulks at the cloud, are not applicable for edge computing, as moving a tremendous amount of data across wide-area networks implies significant transmission delays, not to mention potential privacy leakages during data transmission. However, on-device analytics that requires AI applications to process the IoT data locally may suffer from poor performance and energy efficiency. This is because many AI applications require high computational power that greatly outweighs the capacity of resource- and energy-constrained IoT devices. To address these challenges, this special issue focuses on the overarching architectures, frameworks, and emerging key technologies for the new interdiscipline of edge intelligence.

Digital Object Identifier 10.1109/JIOT.2020.3019948

The response to our call for this special issue was overwhelming, as we received more than 100 submissions from around the world. After a rigorous multiround review process for each article, we were able to accept 35 excellent articles covering various topics in AI-powered edge computing. We would like to express our sincere thanks to Prof. Xuemin (Sherman) Shen (the former Editor-in-Chief), Prof. Honggang Wang (the current Editor-in-Chief), and numerous reviewers for their great support. In the following, we will introduce these articles and highlight their main contributions.

In the article "PreVIous: A methodology for prediction of visual inference performance on IoT devices," Velasco-Montero *et al.* investigated a methodology to predict the performance of convolutional neural networks (CNNs) in terms of throughput and energy consumption on vision-enabled devices for the IoT, in order to select the optimal CNN architecture for a particular hardware platform according to prescribed application requirements.

In the article "Joint DNN partition deployment and resource allocation for delay-sensitive deep learning inference in IoT," He *et al.* studied joint optimization of partition deployment and resource allocation in mobile-edge computing servers to minimize end-to-end inference delay of deep learning tasks in multiple DNN partitions.

In the article "Edge QoE: Computation offloading with deep reinforcement learning for Internet of Things," Lu *et al.* studied the computation offloading service that can offer users with better Quality of Experience (QoE). A double-dueling-deterministic policy gradients (D3PG) algorithm is developed to improve the stability and convergence of deep reinforcement learning (DRL).

In the article "Semisupervised distributed learning with non-IID data for AIoT service platform," Chiu *et al.* considered a novel AIoT (AI + IoT) service platform that collects video data from the individuals' edge devices. An edge learning system based on semi-supervised learning and federated learning technologies is proposed to transform the video into useful information, providing services to IoT applications.

In the article "Stacked autoencoder-based deep-reinforcement learning for online resource scheduling in large-scale MEC networks," Jiang *et al.* proposed a DRL-based solution for minimizing the sum of weighted task latency for all the IoT users, by optimizing offloading decision, transmission power, and resource allocation in the large-scale mobile-edge computing system.

In the article "DAER: A resource preallocation algorithm of edge computing server by using blockchain in intelligent driving," Xiao *et al.* studied the problems of the network economy and resource allocation in edge computing for intelligent driving. A resource transaction architecture based on the blockchain is proposed to eliminate the dependence on third parties, and a dynamic allocation algorithm of edge resources (DAER) based on the double auction mechanism is proposed to maximize the satisfaction of users and service providers of edge computing.

In the article "Dynamic computation offloading with energy harvesting devices: A hybrid-decision-based deep reinforcement learning approach," Zhang *et al.* considered a multidevice–multiserver mobile-edge computing with energy harvesting. Hybrid-decision-based DRL algorithms are proposed for dynamic computation offloading.

In the article "AWARE-CNN: Automated workflow for application-aware real-time edge acceleration of CNNs," Sanchez *et al.* presented AWARE-CNN accelerators for real-time execution of deep learning algorithms on IoT devices. AWARE leverages the reconfigurability of FPGAs to create application-specific architectures customized to match the inherent dataflow of targeted deep neural networks and user-specified real-time requirements.

In the article "Mean field game guided deep reinforcement learning for task placement in cooperative multiaccess edge computing," Shi *et al.* proposed a mean-field game-guided DRL approach for the task placement in the cooperative multiaccess edge computing, which can help servers make timely task placement decisions, and significantly reduce average service delay.

In the article "CEFL: Online admission control, data scheduling, and accuracy tuning for cost-efficient federated learning across edge nodes," Zhou *et al.* investigated how to coordinate the edge and the cloud to optimize the system-wide cost efficiency of federated learning. A cost-efficient optimization framework CEFL based on the Lyapunov optimization theory is developed to make online yet near-optimal control decisions.

In the article "To improve service reliability for AI-powered time-critical services using imperfect transmission in MEC: An experimental study," Liu and Zhang studied the feasibility of UDP-based offloading for AI-powered time-critical services in mobile-edge computing. A prototype is developed and a series of experiments are conducted to understand how image distortion affects inference accuracy.

In the article "Intelligent cooperative edge computing in Internet of Things," Gong *et al.* presented an intelligent cooperative edge computing in IoT networks to achieve a complementary integration of AI and edge computing. A prototype-based evaluation is performed, which indicates that the proposed computing architecture enables a benign combination of AI and edge computing.

In the article "Deep-dual-learning-based cotask processing in multiaccess edge computing systems," Chiang *et al.* investigated the problem of cotask processing in multiaccess edge computing systems. A deep dual learning method is proposed to minimize total cotask completion time.

In the article "Resource optimization for delay-tolerant data in blockchain-enabled IoT with edge computing: A deep reinforcement learning approach," Li *et al.* proposed a joint optimization framework about caching, computation, and security for delay-tolerant data in M2M communication networks based on dueling deep Q-network.

In the article "Machine-learning approach for user association and content placement in fog radio access networks," Yan *et al.* studied the joint user association and cache placement problem in fog radio access networks (F-RANs). DRL-based algorithms are proposed to maximize the F-RAN network payoff.

In the article "Crowd-MECS: A novel crowdsourcing framework for mobile edge caching and sharing," Jiang *et al.* studied the economic and strategic interactions between one content provider and a large crowd of edge devices for crowdsourced mobile-edge caching and sharing. A two-stage Stackelberg game is formulated and analyzed.

In the article "Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching," Wang *et al.* proposed a federated DRL-based cooperative edge caching (FADE) framework. Trace-driven simulation results demonstrate the effectiveness of the proposed FADE framework on reducing the performance loss and average delay, offloading backhaul traffic, and improving the hit rate.

In the article "Energy-efficient processing and robust wireless cooperative transmission for edge inference," Yang *et al.* presented an energy-efficient edge processing framework to execute deep learning inference tasks at the edge computing nodes whose wireless connections to mobile devices are prone to channel uncertainties. A joint inference tasking and downlink beamforming problem is formulated and solved by a statistical learning-based robust optimization approach.

In the article "Swarm-intelligence-based rendezvous selection via edge computing for mobile sensor networks," Liu *et al.* proposed a rendezvous selection strategy for data collection of disjoint wireless sensor networks with mobile edge nodes, in order to achieve full network connectivity and minimize the path length.

In the article "An adaptive dual prediction scheme based on edge intelligence," Wu *et al.* proposed a dual prediction structure based on edge intelligence to maintain the performance of the prediction model with minimum communication cost in long-term prediction. An effective online learning algorithm using adaptive window pattern clustering (AWPC) is proposed to update prediction models online.

In the article "A QoE-aware service-enhancement strategy for edge artificial intelligence applications," Xia *et al.* designed a QoE-aware service enhancement strategy for edge AI applications. Multiple AI algorithms are utilized to execute the same type of tasks concurrently, thus meeting users' heterogeneity requirements of accuracy and delays.

In the article "Deploying network functions for multiaccess edge-IoT with deep reinforcement learning," Shu *et al.* proposed a DRL-based approach for network function deployment. The proposed work can significantly improve the resource utilization of edge servers as well as the computational efficiency of IoT tasks.

In the article "MEC-assisted immersive VR video streaming over terahertz wireless networks: A deep reinforcement learning approach," Du *et al.* minimized the long-term energy consumption of a THz wireless access-based MEC system for high-quality immersive VR video services by jointly optimizing the viewport rendering offloading and downlink transmit power control.

In the article "Personalized federated learning with differential privacy," Hu *et al.* proposed a privacy-preserving approach for learning effective personalized models on distributed user data while guaranteeing the differential privacy of user data. Practical issues in a distributed learning system, such as user heterogeneity, are considered in the proposed approach.

In the article "Lightweight and unobtrusive data obfuscation at IoT edge for remote inference," Xu *et al.* presented a lightweight and unobtrusive approach to obfuscate the inference data at the edge devices. The extensive evaluation shows that the proposed approach effectively protects the confidentiality of the raw forms of the inference data while effectively preserving the backend's inference accuracy.

In the article "FlowGuard: An intelligent edge defense mechanism against IoT DDoS attacks," Jia *et al.* targeted the defense techniques against IoT-oriented DDoS attacks and proposed an edge-centric IoT defense structure (FlowGuard) for the detection, identification, classification, and mitigation of DDoS attacks. A new DDoS attack detection algorithm is presented and two machine learning models for identification and classification based on LSTM and CNN are proposed.

In the article "Topology poisoning attack in SDN-enabled vehicular edge network," Wang *et al.* studied the topology poisoning attack against the SDN controller, implemented this attack in four mainstream controllers, and analyzed its impact.

In the article "Hierarchical incentive mechanism design for federated machine learning in mobile networks," Lim *et al.* proposed a federated learning-based privacy-preserving approach to facilitate collaborative machine learning among multiple model owners in mobile crowdsensing, which allows collaborative machine learning without compromising data privacy given that only the model parameters instead of the raw data are exchanged within the federation.

In the article "*pAElla*: Edge AI-based real-time malware detection in data centers," Libri *et al.* focused on data centers (DCs) and supercomputers (SCs), where a new generation of high-resolution monitoring systems is being deployed. A novel lightweight and scalable approach is proposed to increase the security of DCs/SCs.

In the article "AI at the edge: Blockchain-empowered secure multiparty learning with heterogeneous models," Wang *et al.* proposed a novel blockchain-empowered decentralized secure multiparty learning system with heterogeneous local models called BEMA. Two types of Byzantine attacks are considered, and "off-chain sample mining" and "on-chain mining" schemes are designed to protect the security of the proposed system.

In the article "Mining hard samples globally and efficiently for person reidentification," Sheng *et al.* studied person reidentification (ReID). A new system is introduced for global hard mining to: 1) efficiently mine hard samples from the entire training set and 2) effectively use them in training.

In the article "Modified DenseNet for automatic fabric defect detection with edge computing for minimizing latency," Zhu *et al.* proposed a deep-learning-based fabric defect detection method for edge computing scenarios. A fabric defect detection system is established to test the performance of the optimized model used on edge devices in a real-world textile industry scenario.

In the article "Collaborative data scheduling for vehicular edge computing via deep reinforcement learning," Luo *et al.* studied collaborative data scheduling for vehicular edge computing. A unified framework with communication, computation, caching, and collaborative computing is then formulated, and a collaborative data scheduling scheme to minimize the system-wide data processing cost with ensured delay constraints of applications is developed.

In the article "Edge-computing-enabled unmanned module defect detection and diagnosis system for large-scale photovoltaic plants," Li *et al.* studied module defect detection and diagnosis for large-scale photovoltaic plants. A nondestructive, contactless, and automatical visual inspection system with the help of unmanned aerial vehicles and edge computing is proposed to facilitate defect detection.

In the article "Optimization of edge-PLC-based fault diagnosis with random forest in industrial Internet of Things," Liu *et al.* investigated edge-PLC-based fault diagnosis for industrial IoT and proposed a random forest-based method and an edge-PLCs selection method to save the deployment cost.

Finally, we would like to express our sincere thanks to all the authors for submitting their papers and to the reviewers for their valuable comments and suggestions that significantly enhanced the quality of these articles. We are also grateful to all the editorial staff for their great support throughout the whole review and publication process of this special issue. We hope that this special issue will serve as a useful reference for researchers, scientists, engineers, and academics in the field of AI-powered edge computing for IoT.

LEI YANG, *Guest Editor*
Department of Computer Science
and Engineering
University of Nevada, Reno
Reno, NV 89557 USA

XU CHEN, *Guest Editor*
School of Data and Computer Science
Sun Yat-sen University
Guangzhou 510006, China

SAMIR M. PERLAZA, *Guest Editor*
Centre de Recherche de Sophia Antipolis
—Méditerranée
INRIA
06902 Sophia Antipolis, France

JUNSHAN ZHANG, *Guest Editor*
School of Electrical, Computer,
and Energy Engineering
Arizona State University
Tempe, AZ 85287 USA