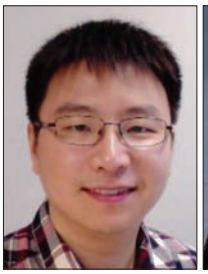
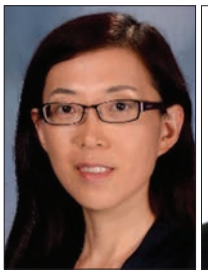
Interplay between Machine Learning and Networking Systems


Xiaowen Chu


Xiaoming Fu


Baochun Li


Wei Wang


Hui Zang


Albert Zomaya

Recently, the advancement of machine learning (ML) techniques, especially deep learning, reinforcement learning, and federated learning, has led to remarkable breakthroughs in a variety of application domains. The success of ML benefits from the advancement of the Internet, mobile networks, data center networks, and the Internet of Things (IoT) that facilitate data creation and sharing. On the other hand, we have also witnessed a fast growing trend in the networking community toward using ML to tackle challenging problems in network design, management, and optimization, which are traditionally addressed using mathematical optimization theory or human-generated heuristics. ML is also an essential ingredient in the realization of autonomous or self-driving networks.

Despite the wide successes of ML-related research in networking systems, there remain many challenges, including the lack of open datasets, open source toolkits, and benchmark suites, reproducibility of the experiments, interpretability and robustness of the ML models, communication bottlenecks in distributed ML systems, and more. The objective of this Special Issue is to bring together the state-of-the-art research results of ML technology and its applications in networking systems. In response to the Call for Papers, we received a total of 59 high-quality submissions from across the world. After a rigorous review process, 15 outstanding papers have been selected for this Special Issue, covering the following five major areas:

1. Network management
2. Mobile and wireless networks
3. Edge and cloud computing
4. Security and privacy
5. ML systems for networking

Area 1: Network Management

In the article "In-Network Neural Networks: Challenges and Opportunities for Innovation," the authors discuss the issues with in-network neural networks that run ML techniques entirely in the forwarding plane. They leverage programmable forwarding planes for managing network events at a nanosecond scale in self-driving networks through in-network intelligence, without performing traffic steering/mirroring to centralized management solutions and minimized human interactions. The authors also present several use cases and new opportunities enabled by distributed in-network intelligence in programmable forwarding planes.

The article "A Brief Survey and Implementation on Refinement for Intent-Driven Networking" presents a generic architecture for intent-driven networking (IDN), and introduces the notion of intent refinement to convert intents from declarative language to a machine-readable policy, which provides a convenient northbound interface for different users to express their communication requirements. To refine different kinds of intent, the authors design an intelligent intent refinement system based on natural language processing (NLP) and deterministic finite automaton (DFA).

In the next article, "Quality Monitoring and Assessment of Deployed Deep Learning Models for Network AIOps," the authors address the challenges of monitoring and assessment of deployed deep learning models in the context of network operation and maintenance. After presenting a broad picture of AIOps, the authors discuss the major challenges of quality assessment, and reviewed the state-of-the-art solutions. With the Open Set Recognition (OSR) techniques identified as a key building block, the authors present two complementary use cases, network traffic classification and image recognition, to verify the feasibility of OSR techniques in model quality assessment.

Area 2: Mobile and Wireless Networks

In the article "Dynamic-Adaptive AI Solutions for Network Slicing Management in Satellite-Integrated B5G Systems," the authors study the challenges of AI-enabled network management and resource orchestration raised by service heterogeneity and rapidly growing network complexity integrated terrestrial and non-terrestrial networks in 5G and beyond 5G. In the context of AI-assisted satellite-B5G network architecture, they examine a set of promising dynamic-adaptive AI solutions for network slicing and case studies, and provide numerical results to illustrate the effectiveness of these AI solutions.

The article "Smarter Base Station Sleeping for Greener Cellular Networks" aims to achieve greener cellular networks by selectively switching some base stations (BSs) with low predicted traffic into sleep mode. To address the challenge of limited traffic log data, the authors leverage the Random Forest algorithm and a feature selection method to predict the traffic using only traffic logs from active BSs. Their evaluation results on a public dataset show the applicability of the proposed solution.

In the article "Predictive Quality of Service: The Next Frontier for Fully Autonomous Systems," the authors present possible ML-based solutions to enable predictive quality of service (PQoS) in autonomous systems. Two important use cases, namely IIoT applications and autonomous driving, are used to illustrate the necessity of PQoS, followed by a review of enabling technologies from the perspective of both 5G core network and radio access network. The authors further present a case study using different ML algorithms to predict the QoS of vehicle-to-everything (V2X) applications.

The next article, "Intimacy-Based Resource Allocation for Network Slicing in 5G via Deep Reinforcement Learning," investigates the resource allocation problem for network slicing in

5G mobile networks. The authors formulate a Markov decision process for the edge controller service request and propose an intimacy-based deep reinforcement learning (DRL) algorithm, I-Slice, to maximize the resource utilization with QoS guarantees. Simulation results show that the proposed I-Slice algorithm outperforms existing methods in terms of resource utilization and delay. The authors also discuss the challenges and some future work for network slicing in 5G networks.

In the article "Machine Learning Assisted Signal Detection in Ambient Backscatter Communication Networks," the authors address the signal detection problem in ambient backscatter communication (AmBC) networks. They first identify the key challenges at different network layers, and then provide a comprehensive survey of ML-based solutions. They further carry out a comparative case study to evaluate the bit error ratio performance of a conventional algorithm and several ML algorithms. Their results show that the deep-learning-based algorithm achieves the best performance due to the strong feature extraction ability. The authors also discuss some open research issues related to signal detection in AmBC networks.

Area 3: Edge and Cloud Computing

In the article "Intelligent Service Orchestration in Edge Cloud Networks," in order to manage software defined networking (SDN)-based transport network and edge cloud (EC) resources, the authors propose a new ML-based data-driven EC selection method for mobile operators, which can dynamically select available ECs even during transport network failures. With emulations using the Graphical Network Simulator-3 (GNS3), the authors show that most of the trained ML models can rather accurately select the correct ECs under the considered two scenarios when transport and EC network parameters are considered in comparison to models trained via only transport or cloud-based parameters. The concept is embedded in an end-to-end mobile network architecture based on extension of the H2020 5Growth1 project's baseline platform.

In the article "Improving Learning-Based DAG Scheduling by Inserting Deliberate Idle Slots," the authors tackle the challenging issue of online job scheduling in large-scale cloud clusters with reinforcement learning (RL). Each online job consists of multiple stages whose precedence constraints are modeled by a directed acyclic graph (DAG). The authors propose to insert deliberate idle time to some jobs so as to reduce the average job completion time (JCT). They develop an RL-based scheduler with carefully designed features for the policy network. The DAGs are encoded into fixed-length vectors by a graph neural network. Experimental results show that inserting deliberate idle time could reduce the average JCT, and the selected features could help to improve the performance of the RL agent.

Area 4: Security and Privacy

The article "NetSpirit: A Smart Collaborative Learning Framework for DDoS Attack Detection" presents a general collaborative learning framework named NetSpirit for effective detection of DDoS attacks. NetSpirit relies on parameter interactions rather than data sharing to mitigate the transmission overhead and protect data privacy. It also integrates semi-supervised learning to make use of unlabeled data. The authors implement a prototype of NetSpirit and evaluate its performance in a simulated environment using a public dataset. The experimental results show that NetSpirit can achieve good detection accuracy with much less network traffic as compared to traditional collaborative learning solutions. They further discuss some open issues and future research directions.

In the article "When Deep Learning Meets Differential Privacy: Privacy, Security, and More," the authors provide a comprehensive review of the recent advances in applying differential privacy (DP) to enhance deep learning (DL) security. They first discuss the essence of DL privacy attacks and DP-based countermeasures. Then they extend their discussion on DP to various aspects, including adversarial attacks, backdoor attacks, fairness-aware DL, reducing overfitting, and interpretability of privacy guarantee. At last, the authors discuss the challenges and future research directions for DP in DL privacy and beyond.

Area 5: ML Systems for Networking

In the article "Decentralized Federated Learning for UAV Networks: Architecture, Challenges, and Opportunities," the authors discuss federated learning (FL) frameworks for unmanned aerial vehicle (UAV) networks. To avoid a single point of failure, they propose a decentralized FL architecture for UAV networks, DFL-UN, in which neighboring UAVs exchange model parameters through device-to-device communications. The authors conduct numerical simulations to evaluate the feasibility and effectiveness of the DFL-UN architecture on the training of a convolutional neural network model. The results show that DFL-UN can achieve similar performance as the centralized FL approach. The authors further discuss the main technical challenges and provide some potential research directions for UAV networks.

The next article, "FEVA: A FEderated Video Analytics Architecture for Networked Smart Cameras," proposes a new federated video analytics architecture called FEVA to address the privacy concern in collaborative video and image analytics. The key idea of FEVA is to keep the video image data local to the edge devices for analytics and only upload the analytics results to the cloud for aggregation. FEVA employs a novel strategy for partitioning video analytics tasks so as to preserve data privacy while maximizing the overall analytics accuracy under the computing and communication constraints of the edge devices. The authors conduct a case study to demonstrate FEVA's performance advantages and privacy-preserving capability in a real-world multi-view 3D vehicles reconstruction task.

In the last article, "IGNNITION: Bridging the Gap between Graph Neural Networks and Networking Systems," the authors propose a novel open source framework called IGNNITION that enables computer network engineers with little neural network programming background to easily develop graph neural network (GNN)-based solutions for many fundamental networking problems, such as topology, routing, and traffic engineering. IGNNITION is based on TensorFlow and provides an intuitive, human-readable YAML interface for users to specify the trained GNN models without coding their implementations. The authors show through case studies that IGNNITION greatly simplifies the training and deployment of variable GNN models in networking systems, while achieving equivalent accuracy and performance for their native implementations in TensorFlow.

In closing, the Guest Editors would like to thank all the authors who submitted their original research work to this Special Issue, and all the reviewers who provided timely and constructive review comments. In addition, we are grateful to Dr. Chonggang Wang, Editor-in-Chief, for his continuous support and guidance, as well as the *IEEE Network* staff for their invaluable support during the course of the preparation of this Special Issue.

Biographies

Xiaowen Chu [SM] received his B.Eng. degree in computer science from Tsinghua University, Beijing, P. R. China, in 1999, and his Ph.D. degree in computer science from the Hong Kong University of Science and Technology (HKUST) in 2003. He was with the Department of Computer Science, Hong Kong Baptist University. He is currently a professor at the Data Science and Analytics Thrust, Information Hub of HKUST (GZ), and an affiliate professor in the Department of Computer Science and Engineering, HKUST. His research interests include GPU computing, distributed machine learning, cloud computing, and wireless networks. He received six Best Paper Awards, including a Best Paper Award of IEEE INFOCOM in 2021. He has served as an Associate Editor or Guest Editor for many journals, including *IEEE Transactions on Network Science and Engineering*, the *IEEE Internet of Things Journal*, *IEEE Network*, and *IEEE Transactions on Industrial Informatics*.

Xiaoming Fu [F] received his Ph.D. from Tsinghua University and is currently a professor of computer science at the University of Göttingen, Germany. He is interested in networked systems and services, cloud computing, mobile computing, big data, and social networks. He has served as a member of several journals' Editorial Boards (e.g., *IEEE TNSM*, *IEEE Communications Magazine*, *Elsevier ComNet*, and *ComCom*) and conference committees (e.g., SIGCOMM, CoNEXT, INFOCOM, ICNP, ICN), and as an elected officer of IEEE ComSoc's Technical Committees on Computer Communications (TCCC) and Internet (ITC).

Baochun Li [F] received his B.Eng. degree from the Department of Computer Science and Technology, Tsinghua University in 1995, and his M.S. and Ph.D. degrees from the Department of Computer Science, University of Illinois at Urbana-Champaign in 1997 and 2000. Since 2000, he has been with the Department of Electrical and Computer Engineering at the University of Toronto, where he is currently a professor. He has held the Bell Canada Endowed Chair in Computer Engineering since August 2005. His research interests include cloud computing, distributed systems, data center networking, and wireless systems. He has co-authored more than 410 research papers, with a total of over 21,000 citations, an H-index of 83, and an i10-index of 280, according to Google Scholar Citations. He was the recipient of the IEEE Communications Society Leonard G. Abraham Award in the Field of Communications Systems in 2000. In 2009, he was a recipient of the Multimedia Communications Best Paper Award from the IEEE Communications Society, and a recipient of the University of Toronto McLean Award. He is a member of ACM.

Wei Wang [M] is an associate professor in the Department of Computer Science and Engineering at HKUST. He is also affiliated with the HKUST Big Data Institute. He received his Ph.D. degree in electrical and computer engineering from the University of Toronto in 2015, and his M.Eng. and B.Eng degrees in electrical engineering from Shanghai Jiao Tong University. His research interests cover the broad area of distributed systems, with focus on data analytics and machine learning systems, cloud computing, and computer networks. He has published extensively in the premier conferences and journals in his fields. His research won the Best Paper Runner Up awards at IEEE ICDCS 2021 and USENIX ICAC 2013. Some of his works have been adopted in Alibaba's production cloud systems. He routinely serves on the Program Committees of leading conferences, and was named as the Distinguished TPC member of IEEE INFOCOM in 2018, 2019, and 2020.

Hui Zang [SM] is currently a tech lead and software engineer at Google. Previously, she was a distinguished scientist at Futurewei Technologies, a lead data scientist at Guavus Inc., and a research scientist at Sprint Labs. She received her B.S. degree in computer science from Tsinghua University, and her M.S. and Ph.D. degrees in computer science from the University of California, Davis. Her research focuses on applying AI and machine learning to build data-driven products and to optimize the performance of computing systems. She is the author of the book *WDM Mesh Networks — Management and Survivability* (Kluwer Academic, 2002) and has published over 70 conference papers and journal articles, with over 30 U.S. patents granted.

Albert Y. Zomaya [F] is a Chair Professor of High-Performance Computing & Networking in the School of Computer Science and director of the Centre for Distributed and High-Performance Computing at the University of Sydney. To date, he has published over 600 scientific papers and articles and is (co-)author/editor of more than 30 books. A sought-after speaker, he has delivered over 250 keynote addresses, invited seminars, and media briefings. His research interests span several areas in parallel and distributed computing and complex systems. He is currently the Editor-in-Chief of *ACM Computing Surveys* and served in the past as Editor-in-Chief of *IEEE Transactions on Computers* (2010–2014) and the Founding Editor-in-Chief of *IEEE Transactions on Sustainable Computing* (2016–2020). He is a decorated scholar with numerous accolades including Fellowship in the American Association for the Advancement of Science and the Institution of Engineering and Technology (United Kingdom). Also, he is an Elected Fellow of the Royal Society of New South Wales, Elected Foreign Member of Academia Europaea, and member of the European Academy of Sciences and Arts. He was the recipient of the 1997 Edgeworth David Medal from the Royal Society of New South Wales for outstanding contributions to Australian science, the IEEE Technical Committee on Parallel Processing Outstanding Service Award (2011), the IEEE Technical Committee on Scalable Computing Medal for Excellence in Scalable Computing (2011), the IEEE Computer Society Technical Achievement Award (2014), the ACM MSWIM Reginald A. Fessenden Award (2017), the New South Wales Premier's Prize of Excellence in Engineering and Information and Communications Technology (2019), and the IEEE Computer Society's Technical Committee on Cloud Computing Research Innovation Award (2021).