PhD Forum Abstract: Intelligence beyond the Edge in IoT

Xiaofan Yu
x1yu@ucsd.edu
University of California San Diego
La Jolla, California, USA

ABSTRACT

Along with the recent advancements of lightweight machine learning and powerful systems and hardware platforms, intelligence beyond the edge has become the next tide of IoT. However, multiple barriers exist from data, algorithm, network and hardware perspectives. In this abstract, I provide an overview of my PhD research which aims at closing the gap towards deploying edge intelligence for large-scale and real-world IoT applications. I further introduce our recent contributions and the work planned ahead.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Sensor networks*; • **Computing methodologies** → **Machine learning**; *Artificial intelligence.*

KEYWORDS

Edge Computing, On-Device Learning, Sensor Networks, IoT

ACM Reference Format:
Xiaofan Yu. 2023. PhD Forum Abstract: Intelligence beyond the Edge in IoT. In *The 22nd International Conference on Information Processing in Sensor Networks (IPSN '23), May 9–12, 2023, San Antonio, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3583120.3589567

1 INTRODUCTION AND CHALLENGES

Along with the recent advancements of lightweight machine learning and powerful hardware platforms (e.g., Raspberry Pi), both training and inference at the edge are possible. In contrast to traditional cloud computing, edge computing first processes the sensory data locally and then transmits the intermediate features to the cloud. Such a paradigm not only reduces the amount of data transmission thus reducing the energy consumption on edge devices, but enables in-situ learning where local algorithms can make decisions in a timely manner. However, multiple **challenges** lie for deploying intelligence beyond the edge for pervasive IoT applications:

- **Data:** The data distribution across *time* may drift due to sensor aging or dynamic environmental changes, while the data collected at separate *locations* can present different distributions, or even different *modalities* if the diverse sensor types are used. Last but not least, supervision in the wilderness is extremely difficult to obtain, if not impossible.

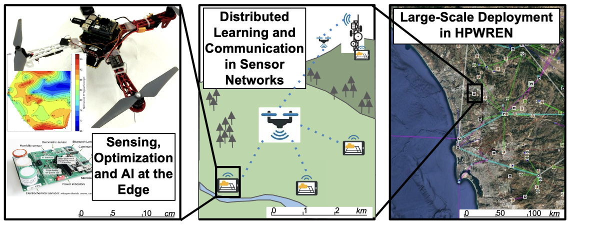
Sensing, Optimization and AI at the Edge
0
5
10
cm
Distributed Learning and Communication in Sensor Networks
0
1
2
km
Large-Scale Deployment in HPWREN
0
50
100
km

Figure 1: Roadmap of my PhD research from single-device AI, distributed AI in sensor networks, to large-scale deployments.

- **Algorithm:** as a result of the data challenges, it is desirable to develop algorithms that are adaptive to drifting and robust to spatial variations under limited supervision. Related fields in ML are continual learning, federated learning, multimodal learning, and self-supervised learning.
- **Network:** in distributed algorithms at the edge (e.g., federated learning), the large models and data to be exchanged present a significant burden when using wireless communications. Joint optimization of learning algorithm and resource allocation is needed in such complex environments.
- **Hardware:** The efficient hardware is a fundamental challenge for all above aspects. How to maximize the benefits of the novel hardware while integrating the whole system for pervasive AI is an active research topic.

My Ph.D. research targets at addressing these barriers and closing the gap towards enabling intelligent in real-world IoT deployments. As shown in Figure 1, I am hoping to contribute a full-stack of technologies from a single device to distributed sensor networks.

2 WORK COMPLETED SO FAR

2.1 Online Self-Supervised Continual Learning

My recent work SCALE [3] focuses on online self-supervised continual learning under (i) sequential data input and (ii) the lack of supervision and prior knowledge at the edge. Consider an autonomous vehicle that learns locally and continually on streaming input during movement (Figure 2). We discover that existing continual learning algorithms rely on supervision or prior knowledge (e.g., the boundary of distribution shifts) for producing good results. To achieve effective continual learning in a practical streaming scenario without task/class labels and prior knowledge, we propose SCALE which extracts and memorizes knowledge on-the-fly as pairwise similarity between representations. SCALE is designed around three major components: (1) a pseudo-supervised contrastive loss, (2) a self-supervised forgetting loss, and (3) an online memory update for uniform subset selection. SCALE outperforms the best state-of-the-art algorithm with improvements up to 3.83%, 2.77% and 5.86% in terms of kNN accuracy on CIFAR-10, CIFAR-100, and SubImageNet datasets.

IPSN '23, May 9–12, 2023, San Antonio, TX, USA

ACM ISBN 979-8-4007-0118-4/23/05.
https://doi.org/10.1145/3583120.3589567

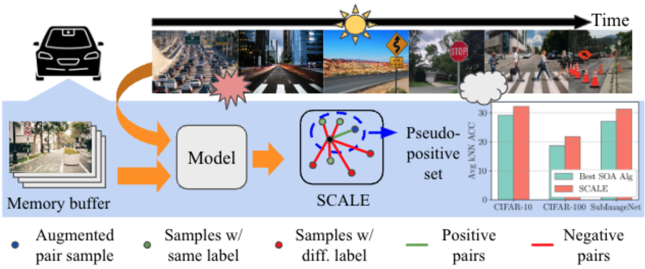

Figure 2: An overview of SCALE [3] that learns by distilling a pseudo-positive set and reinforcing similar samples without supervision and prior knowledge. A memory buffer with carefully designed update is added to prevent catastrophic forgetting.

2.2 Asynchronous Federated Learning

My recent paper on Asynchronous Federated Learning (FL) in hierarchical IoT networks [2] deals with the data and network heterogeneity during deploying FL at the edge. The state-of-the-art FL algorithm (i.e., FedAvg) synchronously aggregates the updated models obtained from local training on edge devices. However, the strong synchrony of FedAvg hinders convergence and robustness if deployed on wireless networks with largely varied delays and unstable connections. We propose an end-to-end framework named Async-HFL for performing FL in a common three-tier IoT network architecture. In response to the networking challenges as shown in Figure 3, Async-HFL employs asynchronous aggregations at both the gateway and the cloud levels thus avoiding long waiting time. To fully unleash the potential of Async-HFL in converging speed under system heterogeneities and stragglers, we design device selection at the gateway level and device-gateway association at the cloud level, both formulated as Integer Linear Program and solved with state-of-the-art optimal solvers. We validate Async-HFL on a physical deployment of 40 clients and observe robust convergence under home Wi-Fi's with largely varied networking speed and unexpected disconnections.

3 WORK PLANNED AHEAD

Moving forward, my ongoing work aims at (1) developing lightweight and robust learning algorithms on edge hardware, and (2) deploying our systems in real-world applications.

3.1 Hyperdimensional Computing

Our team is actively experimenting with new emerging computing paradigms at the edge with custom hardware. Hyperdimensional Computing (HDC) is a novel computing paradigm that is inspired from brain functionality in performing cognitive tasks with high-dimensional sparse representations (a.k.a. hypervectors). HDC is perfect for learning in distributed IoT systems as (i) it supports single-pass training with limited training data, (ii) it is very robust to noise, and (iii) it is highly parallelizable thus enabling hardware implementation. With the emerging ReRAM technologies, HDC achieves magnitudes of energy savings without sacrificing accuracy [1]. One of my ongoing projects is designing an HDC-based algorithm to address in a different way the same online self-supervised continual problem as in the SCALE paper. We plan to achieve a comparable accuracy as traditional neural networks but with faster learning speed and improved energy efficiency by more than 100x on our custom hardware.

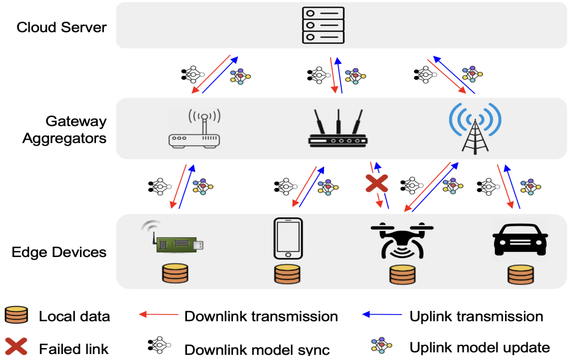

Figure 3: Async-HFL [2] performs asynchronous aggregations at both gateways and cloud to combat heterogeneous delays and unreliable networks in hierarchical IoT networks.

3.2 Large-Scale Real Deployment

Another major thrust is to design the entire framework for large-scale real-world applications. We plan to leverage the High Performance Wireless Research and Education Network (HPWREN), which is an environmental monitoring backbone hosting thousands of sensors and covering an area of 20 k square miles near San Diego County. We plan to set up a testbed as a subset of HPWREN and demonstrate our full-stack technology in real-world ecological applications.

ACKNOWLEDGMENTS

The author would like to thank Prof. Tajana Šimunić Rosing and Dr. Ludmila Cherkasova for their help in revising the manuscript. The research was supported by National Science Foundation under Grants #2112665 (TILOS AI Research Institute), #2003279, #1911095, #1826967, #2100237, #2112167.

REFERENCES

[1] Arpan Dutta et al. 2022. Hdnn-pim: Efficient in memory design of hyperdimensional computing with feature extraction. In *Proceedings of the Great Lakes Symposium on VLSI 2022*. 281–286.

[2] Xiaofan Yu, Ludmila Cherkasova, Harsh Vardhan, Quanling Zhao, Emily Ekaireb, Xiyuan Zhang, Arya Mazumdar, and Tajana Rosing. 2023. Async-HFL: Efficient and Robust Asynchronous Federated Learning in Hierarchical IoT Networks. (2023). To appear in IoTDI'23.

[3] Xiaofan Yu, Yunhui Guo, Sicun Gao, and Tajana Rosing. 2023. SCALE: Online Self-Supervised Lifelong Learning without Prior Knowledge. (2023). To appear in CLVision'23.

A BIOGRAPHY

Xiaofan Yu received the B.S. degree from Peking University, China in 2018 and the M.S. degree from University of California at San Diego in 2020. She is currently pursuing the Ph.D. degree with the Department of CSE, University of California at San Diego. She is working under the supervision of Prof. Tajana Šimunić Rosing, while she has a history of successful collaborations with Dr. Ludmila Cherkasove (Arm Research), Prof. Yunhui Guo (UTDallas), Prof. Arya Mazumdar (UCSD) and Prof. Sicun Gao (UCSD). She is expected to graduate in Spring 2024. She has been invited to participate in the Rising Stars in EECS in 2022.