Third Year Report

Young D. Kwon
Churchill College

UNIVERSITY OF CAMBRIDGE

Efficient Continual Learning and On-Device Training for Mobile and IoT Devices

University of Cambridge
Department of Computer Science and Technology
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom

Email: ydk21@cam.ac.uk

September 30, 2023

Contents

Chapter 1

Progress Updates

Research Statement. My research focuses on building an efficient on-device system that is exceptionally lightweight and capable of updating itself to changing environments and user inputs continually with minimal human intervention.

Third-year Contributions. In my third year, with the increasing need to make tiny MicroController Units (MCUs) intelligent to facilitate various use cases such as smart homes and user customisation, I expand the scope of my research and investigate to what extent I can bring deep learning capability from edge to MCUs (*i.e.* extreme edge).

First, I focus on on-device training at the extreme edge and propose TinyTrain, an on-device training approach that drastically reduces training time by selectively updating parts of the model and explicitly coping with data scarcity. This work has been re-submitted to a top AI conference. Second, I propose LifeLearner, a hardware-aware meta continual learning system that drastically optimises system resources (lower memory, latency, energy consumption) while ensuring high accuracy. This work has been conditionally accepted to SenSys '23 and is currently under shepherding.

Chapter 2

Thesis Outline

The following is the proposed thesis outline:

1. **Introduction.** This chapter introduces the background and motivations to perform continual and on-device learning with real-world application scenarios in mobile computing.

2. **Background.** This chapter describes the relevant research in more details in the areas of on-device ML and CL to discuss the necessity, novelty, and contributions of this thesis.

3. **Initial exploration of continual learning in mobile computing.** This chapter is based on the work [1] published at SEC'21 that investigated the performance and resource trade-offs of various CL methods in mobile sensing tasks of different data modalities. In addition, this chapter discusses the advantages and disadvantages of various CL approaches to provide insights for another work [2] (published at INTERSPEECH'21) that proposed efficient CL by addressing the identified limitations of prior works.

4. **Bringing on-device ML from edge to MCUs: YONO.** This chapter describes the frameworks that explore the interesting area of TinyML designed for extremely resource-constrained platforms, i.e., microcontrollers (MCUs). This chapter explains the first work based on MCUs, YONO [3], published at IPSN'22 which introduced the compression techniques that can support multiple heterogeneous DNNs on MCUs.

5. **Bringing on-device ML from edge to MCUs: TinyTrain.** This chapter describes the second work based on MCUs, TinyTrain [4]. I developed the on-device training framework that jointly leverages data-, memory-, and compute-efficient approach at the extreme edge.

6. **Efficient continual and on-device training on Edge and MCUs.** This chapter explains the final work that orchestrates all the small pieces developed during my PhD to build tiny and efficient CL systems on MCUs [5].

7. **Conclusion.** This chapter will first summarise the overall findings, contributions, and impacts of my research. On top of that, I will discuss the limitations and corresponding future works of this thesis. After that, I will conclude the thesis.

Chapter 3

Timeline

1. Finalise the conditionally accepted work under Shepherding (Michaelmas Term, 2023)
    - Address the reviewers' comments to improve the quality of the paper.
    - Prepare the camera-ready version of the paper (October 2023).
2. Finalise the extension of TinyTrain (Michaelmas Term, 2023)
    - Implement the on-device training approach of TinyTrain on MCUs.
    - Evaluate TinyTrain and other baselines in terms of various aspects such as transfer learning accuracy, latency, energy consumption, and memory and storage usage of on-device training (December 2023).
3. Finalise the thesis writing (Lent Term, 2024)
    - Write the first two chapters of the thesis, i.e., Chapters 1 and 2 (January 2024).
    - Write the rest of the thesis, i.e., Chapters 3, 4, 5, 6, and 7.
    - Prepare the first complete thesis draft with all the chapters.
    - Improve the manuscript by incorporating feedback from my supervisor and colleagues.
    - Finalise the thesis writing (February 2024).
    - Organise the viva and successfully defend the thesis.
    - Finalise the remaining administrative and/or other tasks (March 2024).

Chapter 4

Contributions

In this chapter, I summarise the works related to my PhD thesis. Beyond that, I co-authored some other works in broader areas of mobile systems, machine learning, and human-centred computing.

Papers related to my thesis

[1] *Exploring System Performance of Continual Learning for Mobile and Embedded Sensing Applications*
Young D. Kwon, Jagmohan Chauhan, Abhishek Kumar, Pan Hui, and Cecilia Mascolo.
Proceedings of the Sixth ACM/IEEE Symposium on Edge Computing, 2021. (SEC '21).
Best Paper Award

[2] *FastICARL: Fast Incremental Classifier and Representation Learning with Efficient Budget Allocation in Audio Sensing Applications*
Young D. Kwon, Jagmohan Chauhan, Cecilia Mascolo.
Proceedings of the Annual Conference of the International Speech Communication Association, 2021. (INTERSPEECH '21)

[3] *YONO: Modeling Multiple Heterogeneous Neural Networks on Microcontrollers*
Young D. Kwon, Jagmohan Chauhan, Cecilia Mascolo.
Proceedings of the 21st International Conference on Information Processing in Sensor Networks, 2022. (IPSN '22)

[4] *TinyTrain: Deep Neural Network Training at the Extreme Edge*
Young D. Kwon, Rui Li, Stylianos I. Venieris, Jagmohan Chauhan, Nicholas D. Lane, and Cecilia Mascolo.
Re-submitted to a top AI conference. (Under Review)

[5] *LifeLearner: Hardware-Aware Meta Continual Learning System for Embedded Computing Platforms*
Young D. Kwon, Jagmohan Chauhan, Hong Jia, Stylianos I. Venieris, and Cecilia Mascolo.
Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems, 2023. (SenSys '23). (Under Shepherding)

Other works

UR2M: Uncertainty and Resource-aware Wearable Event Detection on Microcontrollers
Hong Jia, **Young D. Kwon**, Dong Ma, Lorena Qendro, Nhat Pham, Tam Vu, and Cecilia Mascolo.
Re-submitted to a top system conference. (Under Review)

[6] *Exploring User Perspectives on ChatGPT: Applications, Perceptions, and Implications for AI-Integrated Education*
Reza Hadi Mogavi, Chao Deng, Justin Juho Kim, Pengyuan Zhou, **Young D. Kwon**, Ahmed Hosny Saleh Metwally, Ahmed Tlili, Simone Bassanelli, Antonio Bucchiarone, Sujit Gujar, Lennart E Nacke, Pan Hui.
arXiv preprint, 2023.

[7] *Enabling On-Device Smartphone GPU based Training: Lessons Learned*
Anish Das, **Young D. Kwon**, Jagmohan Chauhan, and Cecilia Mascolo.
Proceedings of the 2022 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom '22 Workshops)

[8] *Exploring On-Device Learning Using Few Shots for Audio Classification*
Jagmohan Chauhan, **Young D. Kwon**, and Cecilia Mascolo.
Proceedings of the 30th European Signal Processing Conference, 2022. (EUSIPCO '22)

[9] *PROS: an Efficient Pattern-Driven Operating System for Low-Power Healthcare Wearables*
Nhat Pham, Hong Jia, Minh Tran, Tuan Dinh, Nam Bui, **Young D. Kwon**, Dong Ma, VP Nguyen, Cecilia Mascolo, and Tam Vu.
Proceedings of the 28th Annual International Conference on Mobile Computing and Networking, 2022. (MobiCom '22)

[10] *MyoKey: Inertial Motion Sensing and Gesture-based QWERTY Keyboard for Extended Realities*
Kirill Shatilov, **Young D. Kwon**, Lik-Hang Lee, Dimitris Chatzopoulos, and Pan Hui.
IEEE Transactions on Mobile Computing (TMC), 2022

[11] *Causal Analysis on the Anchor Store Effect in a Location-based Social Network*
Anish Krishna Vallapuram, **Young D. Kwon**, Lik-Hang Lee, Fengli Xu, and Pan Hui.

Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '22).

[12] *Hidenseek: Federated lottery ticket via server-side pruning and sign supermask*
Anish K. Vallapuram, Pengyuan Zhou, **Young D. Kwon**, Lik-Hang Lee, Hengwei Xu, Pan Hui.
arXiv preprint, 2022.

[13] *ContAuth: Continual Learning Framework for Behavioral-based User Authentication*
Jagmohan Chauhan, **Young D. Kwon**, Cecilia Mascolo, and Pan Hui.
Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT/UbiComp), 2021

[14] *Aquilis: Using Contextual Integrity for Privacy Protection on Mobile Devices*
Abhishek Kumar, Tristan BRAUD, **Young D. Kwon**, and Pan Hui.
Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT/UbiComp), 2021

[15] *Interpretable Business Survival Prediction*
Anish Krishna Vallapuram, Nikhil Nanda, **Young D. Kwon**, and Pan Hui.
Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '21).

[16] *IAN: Interpretable Attention Network for Churn Prediction in LBSNs*
Liang-Yu Chen, Yutong Chen, **Young D. Kwon**, Youwen Kang, and Pan Hui.
Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '21).

[17] *Knowing when we do not know: Bayesian continual learning for sensing-based analysis tasks*
Sandra Servia-Rodriguez, Cecilia Mascolo, **Young D. Kwon**.
arXiv preprint, 2021.

Bibliography

[1] Young D Kwon, Jagmohan Chauhan, Abhishek Kumar, Pan Hui, and Cecilia Mascolo. Exploring system performance of continual learning for mobile and embedded sensing applications. In *ACM/IEEE Symposium on Edge Computing*. Association for Computing Machinery (ACM), 2021.

[2] Young D. Kwon, Jagmohan Chauhan, and Cecilia Mascolo. FastICARL: Fast Incremental Classifier and Representation Learning with Efficient Budget Allocation in Audio Sensing Applications. In *Proc. Interspeech 2021*, pages 356–360, 2021.

[3] Young D. Kwon, Jagmohan Chauhan, and Cecilia Mascolo. Yono: Modeling multiple heterogeneous neural networks on microcontrollers. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 285–297, 2022.

[4] Young D. Kwon, Rui Li, Stylianos I. Venieris, Jagmohan Chauhan, Nicholas D. Lane, and Cecilia Mascolo. Tinytrain: Deep neural network training at the extreme edge, 2023.

[5] Young D. Kwon, Jagmohan Chauhan, Hong Jia, Stylianos I. Venieris, and Cecilia Mascolo. Lifelearner: Hardware-aware meta continual learning system for embedded computing platforms. In *21th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2023.

[6] Reza Hadi Mogavi, Chao Deng, Justin Juho Kim, Pengyuan Zhou, Young D. Kwon, Ahmed Hosny Saleh Metwally, Ahmed Tlili, Simone Bassanelli, Antonio Bucchiarone, Sujit Gujar, Lennart E. Nacke, and Pan Hui. Exploring user perspectives on chatgpt: Applications, perceptions, and implications for ai-integrated education, 2023.

[7] Anish Das, Young D. Kwon, Jagmohan Chauhan, and Cecilia Mascolo. Enabling on-device smartphone gpu based training: Lessons learned. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 533–538, 2022.

[8] Jagmohan Chauhan, Young D. Kwon, and Cecilia Mascolo. Exploring on-device learning using few shots for audio classification. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 424–428, 2022.

[9] Nhat Pham, Hong Jia, Minh Tran, Tuan Dinh, Nam Bui, Young Kwon, Dong Ma, Phuc Nguyen, Cecilia Mascolo, and Tam Vu. Pros: An efficient pattern-driven compressive sensing framework for low-power biopotential-based wearables with on-chip intelligence. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, MobiCom '22, page 661–675, New York, NY, USA, 2022. Association for Computing Machinery.

[10] K. Shatilov, Y. D. Kwon, L. Lee, D. Chatzopoulos, and P. Hui. Myokey: Inertial motion sensing and gesture-based qwerty keyboard for extended realities. *IEEE Transactions on Mobile Computing*, (01):1–1, mar 5555.

[11] A. K. Vallapuram, Y. D. Kwon, L. Lee, F. Xu, and P. Hui. Causal analysis on the anchor store effect in a location-based social network. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 202–209, Los Alamitos, CA, USA, nov 2022. IEEE Computer Society.

[12] Anish K. Vallapuram, Pengyuan Zhou, Young D. Kwon, Lik Hang Lee, Hengwei Xu, and Pan Hui. Hidenseek: Federated lottery ticket via server-side pruning and sign supermask, 2022.

[13] Jagmohan Chauhan, Young D. Kwon, Pan Hui, and Cecilia Mascolo. Contauth: Continual learning framework for behavioral-based user authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4), December 2020.

[14] Abhishek Kumar, Tristan Braud, Young D. Kwon, and Pan Hui. Aquilis: Using contextual integrity for privacy protection on mobile devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4), December 2020.

[15] Anish K. Vallapuram, Nikhil Nanda, Young D. Kwon, and Pan Hui. Interpretable business survival prediction. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, pages 99–106, New York, NY, USA, 2021. Association for Computing Machinery.

[16] Liang-yu Chen, Yutong Chen, Young D. Kwon, Youwen Kang, and Pan Hui. Ian: Interpretable attention network for churn prediction in lbsns. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, pages 23–30, New York, NY, USA, 2021. Association for Computing Machinery.

[17] Sandra Servia-Rodriguez, Cecilia Mascolo, and Young D. Kwon. Knowing when we do not know: Bayesian continual learning for sensing-based analysis tasks. *arXiv:2106.05872 [cs]*, June 2021.