Li Yang

PH.D. Candidate
School of Electrical, Computer and Energy Engineering
Arizona State University, USA

+1-407-881-5714
lyang166@asu.edu
linkedin.com/in/LiYang
Scholar.google.com

Research interests

- Efficient and On-device Machine Learning
- Computer Vision
- Transfer/Continual Learning
- Efficient AI Computing Hardware Accelerator Design

Education

2019-current **PH.D., Computer Engineering**
Arizona State University, Tempe, AZ, US

2018-2019 University of Central Florida, Orlando, FL, US
ADVISOR: Deliang Fan

2016-2018 **M.ENG., Electrical Engineering**
University of Central Florida, Orlando, FL, US

2010-2014 **B.S., Information Engineering**
Northeastern University, Qinhuangdao, Hebei, CHN

Professional experience

May-Aug 2022 Applied Scientist intern
Amazon Go Just Walk Out (JWO) team

May-Aug 2021 Deep learning research engineer intern
Samsung Electronics America

2019-current Research/Teaching Assistant
Arizona State University

2016-2019 Research/Teaching Assistant
University of Central Florida

Publications & Presentations (full list)

Research Direction Highlight

- On-device Learning and Efficient Continual Learning:

  CVPR-2022[C1], CVPR-ECV-2022[C2], CVPR-2021[C3], NeurIPS-2022[C4], ICLR-2022[C5],

- Neural Network Model Compression and Dynamic inference:

  NeurIPS-2022[C6], CVPR-2022[C7] TNNLS-2022[J1], DAC-2020 [C8], ASP-DAC-2020 [C9], AAAI-2020 [C10], JETC-2020 [J2]

- Neural Network Accelerator with Hardware and Software Co-design:

  DAC-2022[C11], DATE-2022[C12], ASP-DAC-2022[C13], GLSVLSI-2019 [C14], ISLPED-2018 [C15]

CONFERENCE

[C1] Li Yang, Adnan Siraj Rakin, and Deliang Fan. "Rep-Net: Efficient On-Device Learning via Feature Reprogramming". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

[C2] Li Yang, Adnan Siraj Rakin, and Deliang Fan. "D^A3: Deep Additive Attention Adaption for Memory-Efficient On-Device Multi-Domain Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

[C3] Li Yang, Zhezhi He, Junshan Zhang, and Deliang Fan. "KSM: Fast Multiple Task Adaption via Kernel-wise Soft Mask Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 13845–13853.

[C4] Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. "Beyond Not-Forgetting: Continual Learning with Backward Knowledge Transfer". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[C5] Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. "TRGP: Trust Region Gradient Projection for Continual Learning". In: *International Conference on Learning Representations (ICLR)* (2022).

[C6] Li Yang*, Jian Meng*, Jae-Sun Seo, and Deliang Fan. "Get More at Once: Alternating Sparse Training with Gradient Correction (NeurIPS)". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[C7] Jian Meng, Li Yang, Jinwoo Shin, Deliang Fan, and Jae-sun Seo. "Contrastive Dual Gating: Learning Sparse Features With Contrastive Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 12257–12265.

[C8] Li Yang, Zhezhi He, and Deliang Fan. "Non-uniform DNN Structured Subnets Sampling for Dynamic Inference". In: *57th Design Automation Conference (DAC)* (2020).

[C9] Li Yang, Shaahin Angizi, and Deliang Fan. "A Flexible Processing-in-Memory Accelerator for Dynamic Channel-Adaptive Deep Neural Networks". In: *2020 25rd Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE. 2020.

[C10] Li Yang, Zhezhi He, and Deliang Fan. "Harmonious Coexistence of Structured Weight Pruning and Ternarization for Deep Neural Networks". In: *Thirty-third AAAI Conference on Artificial Intelligence (AAAI)* (2020).

[C11] Fan Zhang, Li Yang, Jian Meng, Jae-sun Seo, Yu Cao, and Deliang Fan. "XMA: a crossbar-aware multi-task adaption framework via shift-based mask learning method". In: *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC)*. 2022, pp. 271–276.

[C12] Fan Zhang, Li Yang, Jian Meng, Jae-Sun Seo, Yu Cao, and Deliang Fan. "XST: A Crossbar Column-wise Sparse Training for Efficient Continual Learning". In: *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE. 2022.

[C13] Fan Zhang, Li Yang, Jian Meng, Yu Kevin Cao, Jae-sun Seo, and Deliang Fan. "Xbm: A crossbar column-wise binary mask learning method for efficient multiple task adaption". In: *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE. 2022.

[C14] Li Yang, Zhezhi He, and Deliang Fan. "Binarized Depthwise Separable Neural Network for Object Tracking in FPGA". In: *Great Lakes Symposium on VLSI (GLVLSI)* (2019).

[C15] Li Yang, Zhezhi He, and Deliang Fan. "A Fully Onchip Binarized Convolutional Neural Network FPGA Implementation with Accurate Inference". In: *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*. ACM. 2018, p. 50.

[C16] Jingbo Sun, Li Yang, Jiaxin Zhang, Frank Liu, Mahantesh Halappanavar, Deliang Fan, and Yu Cao. "Gradient-based novelty detection boosted by self-supervised binary classification". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 36. 8. 2022, pp. 8370–8377.

[C17] Li Yang and Deliang Fan. "Dynamic Neural Network to Enable Run-Time Trade-off between Accuracy and Latency". In: *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE. 2021, pp. 587–592.

[C18] Li Yang, Zhezhi He, Shaahin Angizi, and Deliang Fan. "Processing-in-Memory Accelerator for Dynamic Neural Network with Run-Time Tuning of Accuracy, Power and Latency". In: *2020 IEEE 33rd International System-on-Chip Conference (SOCC)*. IEEE. 2020, pp. 117–122.

JOURNAL

[J1] Li Yang, Zhezhi He, Yu Cao, and Deliang Fan. "A Progressive Subnetwork Searching Framework for Dynamic Inference". In: *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)* (2022).

[J2] Zhezhi He, Li Yang, Shaahin Angizi, Adnan Siraj Rakin, and Deliang Fan. "Sparse BD-Net: a multiplication-less DNN with sparse binarized depth-wise separable convolution". In: *ACM Journal on Emerging Technologies in Computing Systems (JETC)* 16.2 (2020), pp. 1–24.

[J3] Jian Meng, Wonbo Shim, Li Yang, Injune Yeo, Deliang Fan, Shimeng Yu, and Jae-sun Seo. "Temperature-resilient rram-based in-memory computing for dnn inference". In: *IEEE Micro* 42.1 (2021), pp. 89–98.

[J4] Gokul Krishnan, Li Yang, Jingbo Sun, Jubin Hazra, Xiaocong Du, Maximilian Liehr, Zheng Li, Karsten Beckmann, Rajiv V Joshi, Nathaniel C Cady, et al. "Exploring Model Stability of Deep Neural Networks for Reliable RRAM-Based In-Memory Acceleration". In: *IEEE Transactions on Computers (TC)* 71.11 (2022), pp. 2740–2752.

[J5] Jian Meng, Li Yang, Xiaochen Peng, Shimeng Yu, Deliang Fan, and Jae-Sun Seo. "Structured pruning of rram crossbars for efficient in-memory computing acceleration of deep neural networks". In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 68.5 (2021), pp. 1576–1580.

PATENT

[P1] Deliang Fan, Fan Zhang, and Li Yang. *SYSTEM AND METHOD FOR NEURAL NETWORK MULTIPLE TASK ADAPTATION*. United States Patent Application No. 63/369,578 (Patent Pending).

[P2] Li Yang, Adnan Siraj Rakin, and Deliang Fan. *DYNAMIC ADDITIVE ATTENTION ADAPTION FOR MEMORY-EFFICIENT MULTI-DOMAIN ON-DEVICE LEARNING*. United States Patent Application No. 63/351,167 (Patent Pending).

[P3] Deliang Fan, Li Yang, Sen Lin, and Junshan Zhang. *CONTINUAL LEARNING FRAMEWORK THAT GROWS WHEN NECESSARY FOR EFFICIENT LEARNING AT REDUCED COMPLEXITY*. United States Patent Application No. 63/340,305 (Patent Pending).

[P4] Deliang Fan, Adnan Siraj Rakin, Li Yang, Chaitali Chakrabarti, Yu Cao, and Jae-sun Seo. *RA-BNN: Robust & Accurate Binary Neural Network to Simultaneously Defend Adversarial Neural Network Model Noise Injection Attack and Improve Accuracy*. United States Patent Application No. 63/243,762 (Patent Pending).

Awards

2022 DATE-2022 Best Interactive Presentation(IP) Award

2020 DAC-2020 Young Fellow

Hardware Prototype

2018-2019 DAC system design contest 2019: FPGA-based Real-time object tracking via the proposed binary deep neural network (Demo).

Teaching

2018-2019 EEL4758 Computer Architecture: Teaching assistant
2018-2019 EEE3342 Digital System: Teaching assistant
2019-2020 EEE598 Multicore Architecture: Lab instructor
2020-2021 EEE120 Digital Design Fundamentals: Lab instructor

Served as a Reviewer/Program Committee Member

2022 Reviewer of European Conference on Computer Vision (ECCV)
2022 Reviewer of IEEE Transactions on Neural Networks and Learning Systems (TNNLS)
2021 Reviewer of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
2021 Reviewer of International Conference on Machine Learning (ICML)
2020 Reviewer of Design Automation Conference (DAC)
2019 Reviewer of ACM Great Lakes Symposium on VLSI (GLSVLSI)