Pu Zhao

Cell: (+1) 617-606-2116 Email: {zhao.pu, p.zhao}@northeastern.edu Google Scholar

RESEARCH INTERESTS

- Key words: DNN efficiency, model pruning, sparsity, neural architecture search, real-time inference, edge device, DNN robustness, adversarial attacks, super resolution, segmentation, 3D detection, optimization, DNN fingerprinting, DNN IP protection, efficient transformers, compiler optimization
- Efficient deep learning
  - Investigate DNN pruning methods to obtain sparse DNN models with less computations & parameters, and comparable accuracy performance, thus improving DNN model efficiency
  - Design neural architecture search methods to automatically search small DNN models with high accuracy and fast inference
  - Achieve real-time inference for DNN models with compiler optimization on resource-limited hardware devices such as mobile phones, thus facilitating DNN deployment on edge devices
  - Improve the model efficiency of various tasks such as super resolution, segmentation, 3D detection and so on, which usually cost huge computations and memory
- DNN Robustness
  - Explore various DNN attack methods such as adversarial attack, fault injection attack and so on
  - Investigate defense methods against various DNN attacks
- DNN Fingerprinting and IP protection
  - Propose intrinsic examples for DNN fingerprinting to verify the functionality of DNN models, specifically, detect adversarial third-party attacks while enabling normal model transformations
- Optimization methods
  - Various optimization methods are adopted including: bi-level optimization, Alternating Direction Method of Multipliers, Bayesian optimization, natural gradient descent and Fisher information, second-order optimization, zeroth-order optimization, min-max optimization and so on

EDUCATION BACKGROUND

Ph.D. in Electrical and Computer Engineering, Northeastern University Sept. 2016-Aug. 2021
- Major: Computer Engineering (GPA: 3.75/4)
- Advisor: Prof. Xue Lin **(**https://web.northeastern.edu/xuelin/**)**

M.E. in Electrical Engineering, Shanghai Jiao Tong University Sept. 2013-Mar. 2016
- Major: Electronic and Communication Engineering (GPA: 2.59/3.0)

B.S. in Electrical Engineering, Shanghai Jiao Tong University Sept. 2009-Jun. 2013
- Major: Information Engineering (GPA: 86/100)

POSITIONS

Research Assistant Professor in Electrical and Computer Engineering **Northeastern University**
- Improve DNN efficiency of various tasks (such as super-resolution and segmentation) to achieve real-time inference on resource-limited hardware (such as mobile phones) Sept. 2021-Present

Research Internship **Google**
- Explore image retrieval based on a reference image and a modification text May 2020-Aug. 2020

Research Fall Internship-Graduate **IBM Thomas J. Watson Research Center**
- Improve DNN robustness under various attacks with model connection Dec. 2018-Mar. 2019

PUBLICATIONS

(* equal contribution)

Conference Papers

1. [**AAAI'23**] Yanyu Li*, Changdi Yang*, Pu Zhao*, Geng Yuan, Wei Niu, Jiexiong Guan, Hao Tang, Minghai Qin, Qing Jin, Bin Ren, Xue Lin, and Yanzhi Wang, Towards Real-Time Segmentation on the Edge, In *Association for the Advancement of Artificial Intelligence*, 2023.
2. [**AAAIW'23**] Yize Li, Pu Zhao, Xue Lin, Bhavya Kailkhura, and Ryan Goldhahn, Less is More: Data Pruning for Faster Adversarial Training, In *SafeAI, The AAAI's Workshop on Artificial Intelligence Safety*, 2023.
3. [**NeurIPS'22**] Yihua Zhang, Yuguang Yao, Parikshit Ram, Pu Zhao, Tianlong Chen, Mingyi Hong, Yanzhi Wang, and Sijia Liu, Advancing Model Pruning via Bi-level Optimization, In *Neural Information Processing Systems*, 2022. **Acceptance rate: 25.6% (2665/10,411)**
4. [**ICCAD'22**] Yifan Gong, Zheng Zhan, Pu Zhao, Yushu Wu, Chao Wu, Caiwen Ding, Weiwen Jiang, Minghai Qin, and Yanzhi Wang. All-in-One: A Highly Representative DNN Pruning Framework for Edge Devices with Dynamic Power Management. In *International Conference on Computer Aided Design*, ACM, 2022. **Acceptance rate: 22.5% (132/586)**
5. [**ECCV'22**] Yushu Wu, Yifan Gong, Pu Zhao, Yanyu Li, Zheng Zhan, Wei Niu, Hao Tang, Minghai Qin, Bin Ren, and Yanzhi Wang. Compiler-Aware Neural Architecture Search for On-Mobile Real-time Super-Resolution. In *European Conference on Computer Vision*, 2022. **Acceptance rate: 28% (1650/5803)**
6. [**IJCAI'22**] Pu Zhao, Parikshit Ram, Songtao Lu, Yuguang Yao, Djallel Bouneffouf, Xue Lin, and Sijia Liu. Learning to Generate Image Source-Agnostic Universal Adversarial Perturbations. In *International Joint Conferences on Artificial Intelligence Organization*, 2022. **Acceptance rate: 15%**
7. [**IJCAI'22**] Yanyu Li, Pu Zhao, Geng Yuan, Xue Lin, Yanzhi Wang and Xin Chen. Pruning-as-Search: Efficient Neural Architecture Search via Channel Pruning and Structural Reparameterization. In *International Joint Conferences on Artificial Intelligence Organization*, 2022. **Acceptance rate: 15%**
8. [**WACVW'22**] Hao Cheng, Kaidi Xu, Zhengang Li, Pu Zhao, Chenan Wang, Xue Lin, Bhavya Kailkhura, and Ryan Goldhahn. More or Less (MoL): Defending against Multiple Perturbation Attacks on Deep Neural Networks through Model Ensemble and Compression. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2022
9. [**ISQED'22**] Xiaolong Ma*, Geng Yuan*, Zhengang Li*, Yifan Gong, Tianyun Zhang, Wei Niu, Zheng Zhan, Pu Zhao, Jian Tang, Xue Lin, Bin Ren, and Yanzhi Wang. BLCR: towards real-time DNN execution with block-based reweighed pruning. In *Proceedings of the 23rd International Symposium on Quality Electronic Design*, IEEE, 2022. **Acceptance rate: 30%**
10. [**ICCV'21**] Zheng Zhan, Yifan Gong, Pu Zhao, Geng Yuan, Wei Niu, Yushu Wu, Tianyun Zhang, Malith Jayaweera, David Kaeli, Bin Ren, Xue Lin, and Yanzhi Wang. Achieving on-Mobile Real-Time Super Resolution with Neural Architecture and Pruning Search. In *International Conference on Computer Vision*, 2021. **Acceptance rate: 25.9% (1617/6236)**
11. [**IJCAI'21**] Siyue Wang, Xiao Wang, Pin-Yu Chen, Pu Zhao, and Xue Lin. Characteristic Examples: High-Robustness, Low-Transferability Fingerprinting of Neural Networks. In *International Joint Conferences on Artificial Intelligence Organization*, 2021. **Acceptance rate: 13.9% (587/4204)**
12. [**ICLR'21 Workshop**] Siyue Wang, Xiao Wang, Pin-Yu Chen, Pu Zhao, and Xue Lin. Characteristic examples: high-robustness, low-transferability fingerprinting of neural networks. In *International Conferences on Learning Representations Workshop on Security and Safety in Machine Learning Systems*, 2021.
13. [**BMVC'21**] Siyue Wang, Pu Zhao, Xiao Wang, Sang Chin, Thomas Wahl, Yunsi Fei, Qi Chen, and

Xue Lin. Intrinsic Examples: Robust Fingerprinting of Deep Neural Networks. In *The 32nd British Machine Vision Conference,* 2021. **Acceptance rate: 36% (437/1206)**

14. [**CVPR'21**][**Oral (top 5%)**] Zhengang Li*, Pu Zhao*, Geng Yuan*, Wei Niu*, Yanyu Li, Yuxuan Cai, Xuan Shen, Zheng Zhan, Zhenglun Kong, Qing Jin, Zhiyu Chen, Sijia Liu, Kaiyuan Yang, Bin Ren, Yanzhi Wang, and Xue Lin, NPAS: A Compiler-aware Framework of Unified Network Pruning and Architecture Search for Beyond Real-Time Mobile Acceleration. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2021*. **Acceptance rate: 25%**
15. **[RTAS'21 Brief Presentation]** Pu Zhao, Wei Niu, Geng Yuan, Yuxuan Cai, Hsin-Hsuan Sung, Shaoshan Liu, Sijia Liu, Xipeng Shen, Bin Ren, Yanzhi Wang, and Xue Lin. Brief Industry Paper: Towards real-time 3D object detection for autonomous vehicles with pruning search. In *Proceedings of the 27th IEEE Real-Time and Embedded Technology and Applications Symposium*, 2021.
16. [**DAC'21**] Pu Zhao, Geng Yuan, Yuxuan Cai, Wei Niu, Qi Liu, Wujie Wen, Bin Ren, Yanzhi Wang, Xue Lin, Neural Pruning Search for Real-Time Object Detection of Autonomous Vehicles. In *Proceedings of the 58th Annual Design Automation Conference 2021*. **Acceptance rate: 23%**
17. [**AdvML'21**] Yize Li, Pu Zhao, Yuguang Yao, Vishal Asnani, Yifan Gong, Yimeng Zhang, Zhengang Li, Xiaoming Liu, Sijia Liu, and Xue Lin. Supervised classification on deep neural network attack toolchains. In *the 3rd Workshop on Adversarial Learning Methods for Machine Learning and Data Mining at KDD*, 2021.
18. [**AdvML'21**] Chenan Wang, Pu Zhao, Siyue Wang, and Xue Lin. Detection and recovery against deep neural network fault injection attacks based on contrastive learning. In *the 3rd Workshop on Adversarial Learning Methods for Machine Learning and Data Mining at KDD*, 2021.
19. **[CogArch'21]** Pu Zhao, Wei Niu, Geng Yuan, Yuxuan Cai, Bin Ren, Yanzhi Wang, and Xue Lin. Achieving real-time object detection on mobile devices with neural pruning search. Presented at *the 5th Workshop on Cognitive Architectures in HPCA*, 2021.
20. **[AccML'21]** Pu Zhao, Geng Yuan, Yuxuan Cai, Wei Niu, Bin Ren, Yanzhi Wang, and Xue Lin. Neural pruning search for real-time object detection of autonomous vehicles. Presented at *the 3rd Accelerated Machine Learning Workshop in HiPEAC,* 2021.
21. **[AccML'21]** Zhengang Li, Geng Yuan, Wei Niu, Yanyu Li, Pu Zhao, Yuxuan Cai, Xuan Shen, Zheng Zhan, Zhenglun Kong, Qing Jin, Bin Ren, Yanzhi Wang, and Xue Lin. NPS: a compiler-aware framework of unified network pruning for beyond real-time mobile acceleration. Presented at *the 3rd Accelerated Machine Learning Workshop in HiPEAC,* 2021.
22. [**IJCAI-PRICAI'20**] Wei Niu*, Pu Zhao*, Zheng Zhan, Xue Lin, Yanzhi Wang and Bin Ren. Towards Real-Time DNN Inference on Mobile Platforms with Model Pruning and Compiler Optimization. In IJCAI-PRICAI, 2020.
23. **[ISLPED'20 Design Contest 1st Place]** Geng Yuan, Wei Niu, Pu Zhao, Xue Lin, Bin Ren, and Yanzhi Wang. CoCoPIE: A framework of compression-compilation co-design towards ultra-high energy efficiency and real-time DNN inference on mobile devices. In *ACM/IEEE International Symposium on Low Power Electronics and Design*, 2020.
24. **[ECCV'20 Demo]** Wei Niu, Mengshu Sun, Zhengang Li, Geng Yuan, Pu Zhao, Xue Lin, and Bin Ren. Real-time 3D CNN Inference for Action Recognition on Mobile Devices. In *European Conference on Computer Vision*, 2020.
25. **[ECCV'20 Demo]** Zheng Zhan, Pu Zhao, Geng Yuan, Wei Liu, Bin Ren, and Xue Lin. Realtime DNN inferences on mobile devices for various practical deep learning applications. In *European Conference on Computer Vision*, 2020.
26. [**DAC'20**] Mengshu Sun*, Pu Zhao*, Mehmet Gungor, Massoud Pedram, Miriam Leeser, Xue Lin. 3D CNN Acceleration on FPGA using Hardware-Aware Pruning. In *Proceedings of the 57th Annual Design Automation Conference 2020.* **Acceptance rate: 21% (158/ 747)**
27. [**ICLR'20**] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, Xue Lin. Bridging Mode Connectivity in Loss Landscapes and Adversarial Robustness. In *International Conferences on*

Learning Representations 2020. **Acceptance rate: 26.5% (687/2594)**

28. [**AAAI'20**] Pu Zhao, Pin-Yu Chen, Siyue Wang, and Xue Lin. Towards query-efficient black-box adversary with zeroth-order natural gradient descent. In *Association for the Advancement of Artificial Intelligence, 2020.* **Acceptance rate: 20.6% (1591/7737)**
29. [**AAAI'20**] Lily Weng*, Pu Zhao*, Sijia Liu, Pin-Yu Chen, Xue Lin, and Luca Daniel. Towards certificated model robustness against weight perturbations. In *Association for the Advancement of Artificial Intelligence, 2020.* **Acceptance rate: 20.6% (1591/7737)**
30. [**ICCV'19**] Pu Zhao, Sijia Liu, Pin-Yu Chen, Nghia Hoang, Kaidi Xu, Bhavya Kailkhura, Xue Lin. On the Design of Black-box Adversarial Examples by Leveraging Gradient-free Optimization and Operator Splitting Method. In *International Conference on Computer Vision.* **Acceptance rate: 25% (1077/4303)**
31. [**DAC'19**] Pu Zhao, Siyue Wang, Cheng Gongye, Yanzhi Wang, Yunsi Fei, and Xue Lin. Fault Sneaking Attack: a Stealthy Framework for Misleading Deep Neural Networks. In *Proceedings of the 56th Annual Design Automation Conference 2019.* **Acceptance rate: 24.8% (202/815)**
32. [**GLSVLSI'19**] Mengshu Sun, Pu Zhao, Yanzhi Wang, Naehyuck Chang, and Xue Lin. HSIM-DNN: Hardware Simulator for Computation-, Storage- and Power-Efficient Deep Neural Networks. In *Proceedings of Great Lakes Symposium on VLSI*, ACM, 2019. **Acceptance rate: 29%**
33. [**ICLR'19**] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured Adversarial Attack: Towards General Implementation and Better Interpretability. In *International Conferences on Learning Representations, 2019.* **Acceptance rate: 31% (500/1591)**
34. [**ASP-DAC'19**] Pu Zhao, Kaidi Xu, Sijia Liu, Yanzhi Wang, and Xue Lin. Admm attack: an enhanced adversarial attack for deep neural networks with undetectable distortions. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, pages 499-505. ACM, 2019.
35. [**GlobalSIP'18**] Pu Zhao, Kaidi Xu, Tianyun Zhang, Makan Fardad, Yanzhi Wang and Xue Lin. Reinforced Adversarial Attacks on Deep Neural Networks Using ADMM. In *2018 IEEE Global Signal Processing for Adversarial Machine Learning*.
36. [**ICCAD'18**][**Best Paper Nomination**] Siyue Wang, Xiao Wang, Pu Zhao, Wujie Wen, David Kaeli, Peter Chin and Xue Lin. Defensive Dropout for Hardening Deep Neural Networks under Adversarial Attacks. In *2018 International Conference on Computer Aided Design*, ACM, 2018. **Acceptance rate: 25% (98/396)**
37. [**ACM MM'18**] Pu Zhao, Sijia Liu, Yanzhi Wang and Xue Lin. An ADMM-Based Universal Framework for Adversarial Attacks on Deep Neural Networks. In *2018 ACM Multimedia Conference (MM'18)*, ACM, 2018. **Acceptance rate: 27.5% (144/757)**
38. [**ASP-DAC'18**] Pu Zhao, Yanzhi Wang, Naehyuck Chang, Qi Zhu and Xue Lin. A Deep Reinforcement Learning Framework for Optimizing Fuel Economy of Hybrid Electric Vehicles. In *2018 23rd Asia and South Pacific Design Automation Conference*, pages 196-202. IEEE, 2018.
39. [**WCSP'14**] Pu Zhao, M. Zhang, H. Yu, H. Luo, and W. Chen, Beamforming design of sum secrecy rate optimization for MU-MISO channel under sum power constraint, in *the 6th International Conference on Wireless Communications and Signal Processing*, 2014.
40. [**CHINACOM'13**] Mengshi Li, Yisha Lou, Pu Zhao, Hui Yu, and Xiao Tian, Correlation and capacity evaluation for dual polarized MIMO system based on an extended 3-D geometry-based stochastic channel model, in *8th International Conference on Communications and Networking in China*. 2013.

Journal Papers

41. [**CACM'20**][**Impact Factor: 4.6**] Hui Guan, Shaoshan Liu, Xiaolong Ma, Wei Niu, Bin Ren, Xipeng Shen, Yanzhi Wang, Pu Zhao. CoCoPIE: Enabling Real-Time AI on Off-the-Shelf Mobile Devices via

Compression-Compilation Co-Design. In *Communications of the ACM*, Vol. 64 No. 6, Pages 62-68. 10.1145/3418297, June 2021

42. [**Parallel Computing'20**][**Impact Factor: 1.12**] Shi Dong, Pu Zhao, Xue Lin, and David Kaeli. Exploring GPU acceleration of deep neural networks using block circulant matrices. https://doi.org/10.1016/j.parco.2020.102701, Volume 100, December 2020
43. [**IET-CPS'17**][**Impact Factor: 2.7**] Pu Zhao, Xue Lin, Yanzhi Wang, Shuang Chen and Massoud Pedram. A Hierarchical Resource Allocation and Consolidation Framework in a Multi-Core Server Cluster Using a Markov Decision Process Model. In *IET Cyber-Physical Systems: Theory & Applications*, vol. 2, no. 3, pp. 118-126, October 2017.
44. [**T IFS'15**][**Impact Factor: 7.2**] Pu Zhao, Meng Zhang, Hui Yu, Hanwen Luo and Wen Chen. Robust Beamforming Design for Sum Secrecy Rate Optimization in MU-MISO Networks. In *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 9, pp. 1812–1823, Apr. 2015.

SYNERGISTIC ACTIVITIES

Conference Technical Program Committee Member:

- Design Automation Conference (DAC), 2022~2023
- European Conference on Computer Vision (ECCV), 2022
- International Conference on Machine Learning (ICML), 2021
- International Conference on Computer Vision (ICCV), 2021
- International Conferences on Learning Representations (ICLR), 2021~2022
- Neural Information Processing Systems (NeurIPS), 2020-2022
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020-2023
- National Conference on Artificial Intelligence (AAAI), 2020-2023

Journal Reviewer:

- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2020-2021
- IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 2020-2021
- Transactions on Machine Learning Research (TMLR), 2022
- ACM Computing Surveys, 2020-2021
- Signal Processing: Image Communication, 2022
- Cybersecurity, 2022

HONORS AND AWARDS

2022	Work with Prof. Lin and win **U.S. DOT Inclusive Design Challenge Stage II** together with VEMI Lab from University Maine
2020	**Design Contest 1st Place** on 2020 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)
2018	**ICCAD 2018 Best Paper Candidate**
2014, 2015	**National Scholarship** for Postgraduate Students at Shanghai Jiao Tong University (5/180)
2013, 2014	**Academic Excellence Scholarship** for Postgraduate Students at Shanghai Jiao Tong University
2011, 2012	**Academic Excellence Scholarship** of Shanghai Jiao Tong University
2009	26th National Physics Contest for College Students, **First Prize**

SKILLS

- Programming Languages: C/C++, Python, MATLAB, Verilog
- Deep Learning: TensorFlow, PyTorch, Keras

ADDITIONAL EXPERIENCE

2018, 2020 **Teaching Assistant** for Computer Architecture — Northeastern University