SHENG LI

✉ shl188@pitt.edu 🏠 Homepage in LinkedIn 🎓 Google Scholar

EDUCATION

University of Pittsburgh *Aug. 2021 ∼ Present*
Pittsburgh, PA, USA
Ph.D. in Computer Science
(with a focus on Efficient Machine Learning Algorithms and Systems)
Advisor: Dr. Xulong Tang

Sichuan University *Sept. 2016 ∼ June 2020*
Chengdu, Sichuan, China
B.Eng. in Software Engineering, Graduated with Honor
Advisor: Dr. Ziliang Feng

PUBLICATIONS

**equal contribution*

[1] **Sheng Li***, Qitao Tan*, Yue Dai, Zhenglun Kong, Tianyu Wang, Jun Liu, Ao Li, Ninghao Liu, Yufei Ding, Xulong Tang, and Geng Yuan. "Mutual Effort for Efficiency: A Similarity-based Token Pruning for Vision Transformers in Self-Supervised Learning." *In Proceedings of 13th International Conference on Learning Representations.* **(ICLR 2025)**

[2] Huidong Ji, **Sheng Li**, Yue Cao, Chen Ding, Jiawei Xu, Qitao Tan, Ao Li, Jun Liu, Xulong Tang, Lirong Zheng, Geng Yuan, and Zhuo Zou. "A Computation and Energy Efficient Hardware Architecture for SSL Acceleration." *In Proceedings of 30th Asia and South Pacific Design Automation Conference.* **(ASP-DAC 2025)**

[3] **Sheng Li**, Chao Wu, Ao Li, Yanzhi Wang, Xulong Tang, and Geng Yuan. "Waxing-and-Waning: a Generic Similarity-based Framework for Efficient Self-Supervised Learning." *In Proceedings of 12th International Conference on Learning Representations.* **(ICLR 2024)**

[4] **Sheng Li***, Geng Yuan*, Yue Dai*, Youtao Zhang, Yanzhi Wang, and Xulong Tang. "SmartFRZ: An Efficient Training Framework using Attention-Based Layer Freezing." *In Proceedings of 11th International Conference on Learning Representations.* **(ICLR 2023, Honored with Spotlight Award)**

[5] **Sheng Li**, Geng Yuan, Yawen Wu, Yue Dai, Tianyu Wang, Chao Wu, Alex K. Jones, Jingtong Hu, Yanzhi Wang, and Xulong Tang. "ETuner: A Redundancy-Aware Framework for Efficient Continual Learning Application on Edge Devices." **(Preprint)**

[6] Tianyu Wang*, **Sheng Li***, Bingyao Li, Yue Dai, Ao Li, Geng Yuan, Yufei Ding, Youtao Zhang, and Xulong Tang. "Improving GPU Multi-Tenancy Through Dynamic Multi-Instance GPU Reconfiguration." **(Preprint)**

[7] Geng Yuan*, Yanyu Li*, **Sheng Li**, Zhenglun Kong, Sergey Tulyakov, Xulong Tang, Yanzhi Wang, and Jian Ren. "Layer Freezing & Data Sieving: Missing Pieces of a Generic Framework for Sparse Training". *In Proceedings of the 36th Conference on Neural Information Processing Systems.* **(NeurIPS 2022)**

[8] Sébastien Ollivier, **Sheng Li**, Yue Tang, Stephen Cahoon, Ryan Caginalp, Chayanika Chaudhuri, Peipei Zhou, Xulong Tang, Jingtong Hu, and Alex K. Jones. "Sustainable AI Processing at the Edge." *IEEE Micro, 2022.* **(IEEE Micro)**

[9] Bingyao Li*, Qi Xue*, Geng Yuan*, **Sheng Li**, Xiaolong Ma, Yanzhi Wang, and Xulong Tang. "Optimizing Data Layout for Training Deep Neural Networks." *In Companion Proceedings of the Web Conference 2022.* **(WWW 2022 workshop)**

[10] Mingzheng Hou, Ziliang Feng, Haobo Wang, Zhiwei Shen, and **Sheng Li**. "An adaptive regression based single-image super-resolution." *Multimedia Tools and Applications, 2022.* **(Multimed. Tools Appl.)**

[11] **Sheng Li**, Zanhan Ding, and Honglv Chen. "A Neural Network-Based Teaching Style Analysis Model." *In 2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics, IEEE, 2019.* **(IHMSC 2019)**

[12] **Sheng Li**, and Hanxin Feng. "EEG signal classification method based on feature priority analysis and CNN." *In 2019 International Conference on Communications, Information System and Computer Engineering, IEEE, 2019.* **(CISCE 2019)**

RESEARCH EXPERIENCE

University of Pittsburgh *2020 - Present*

Research Assistant / Teaching Assistant

Advisor: Dr. Xulong Tang

- We propose an efficient layer freezing approach that can reduce computation cost by 40% and memory costs by 50% for ML model training, without affecting the model performance.
- We propose an efficient self-supervised learning technique that prunes less important parts of input image, cutting training costs by over 40% without sacrificing model performance.
- We propose a token pruning approach for vision transformers in self-supervised learning, reducing the training costs by 40% without influencing the resultant model performance.
- We propose a framework to reduce the training redundancy in continual learning on resource-constrained edge devices. Our framework reduces the computation cost and energy by more than 80% without compromising the model accuracy.

Sichuan Universiity *2019*

Research Assistant

Advisor: Dr. Ziliang Feng

- We apply machine learning algorithms to real-world applications such as quality inspection of building materials in civil engineering.
- We design a fatigue-driving alert system based on driver brainwave detection and analysis.

SELECTED HONORS AND AWARDS

Spotlight Award, The 11th ICLR Conference	*2023*
National Scholarship, Ministry of Education of China	*2018, 2019*
Comprehensive Scholarship, Sichuan University	*2017*

TEACHING

2022	Teaching Assistant	CS 1501: Algorithms and Data Structures
2021, 2022	Teaching Assistant	CS 0447: Computer Organization and Assembly Language

PROFESSIONAL SERVICE

Reviewer of NeurIPS, ICLR, ICML, ACML, and Neural Processing Letters (NPL)