Huanrui Yang

Assistant Professor, ECE, The University of Arizona
1230 E. Speedway Blvd., Tucson, AZ 85721
Cell: 919-638-7799 | E-mail: huanruiyang@arizona.edu | Homepage: https://sites.google.com/view/huanrui-yang
Updated in June 2024

RESEARCH INTEREST

- Develop mathematical understandings of the efficiency and robustness of compound AI systems.
- Explore new learning and evaluation schemes for better generalizability, robustness, and interpretability.
- Deep learning privacy, interpretability, federated learning, and software-hardware co-design.

PROFESSIONAL EXPERIENCE

The University of Arizona 08/2024—now
Assistant Professor, Department of Electrical & Computer Engineering

TetraMem Inc. 05/2024—08/2024
Visiting Scholar

University of California, Berkeley 06/2022—05/2024
Postdoctoral Scholar

- Supervised by Prof. Kurt Keutzer.
- Research on efficient deep learning for computer vision, speech recognition and natural language processing.
- Mentoring PhD students, MENG, and 5th Year Master students.

NVIDIA Corporation 02/2021—09/2021
Research Intern

- Supervised by Dr. Danny Yin, Dr. Pavlo Molchanov and Dr. Jan Kautz.
- Research on Vision Transformer compression and efficient parameter redistribution rules.

Microsoft Corporation 05/2018—08/2018
Research Intern

- Supervised by Dr. Wenhan Wang and Dr. Yuxiong He.
- Research on model compression technique for large RNN/LSTM using SVD decomposition.

EDUCATION

Duke University 08/2017 to 05/2022
Ph.D., Electrical and Computer Engineering
Dissertation Title: Towards Efficient and Robust Deep Neural Network Models
Advisor: Prof. Hai Li and Prof. Yiran Chen

- Duke ECE Outstanding Service Award

Tsinghua University 08/2013 to 07/2017
B.E., Electronic Engineering
Diploma Thesis Title: On-chip Trainable Fully Connected Neural Network Accelerator Architecture
Advisor: Prof. Yongpan Liu

- Outstanding Diploma Thesis of Tsinghua University

Experimental Class for Gifted Children, Beijing No.8 Middle School 09/2009 to 06/2013
High School

- Outstanding Graduates of Beijing No.8 Middle School (Top 10)

TEACHING EXPERIENCE

ECE 550D Fundamentals of Computer Systems and Engineering, Duke University Fall 2018
Teaching Assistant
Instructor: Prof. Hai Li and Prof. Yiran Chen

ECE 681 Pattern Classification and Recognition, Duke University Spring 2019
Teaching Assistant
Instructor: Prof. Stacy Tantum

ECE 661 Computer Engineering Machine Learning and Deep Neural Nets, Duke University
Leading Teaching Assistant and Substitute Instructor Fall 2019, Fall 2020 & Fall 2021
Instructor: Prof. Hai Li and Prof. Yiran Chen

AWARDS AND HONORS

2021 Shanghai World AI Conference Yunfan Award (Future star)	07/2021
Duke ECE Outstanding Service Award	05/2022
Outstanding Diploma Thesis, Tsinghua University	06/2017
Best paper runner-up in IEEE/ACM CHASE 2024	06/2024
Best paper award in MICRO 2021	11/2021
Oral presentation in NeurIPS 2020 (Top 1%)	12/2020
Best student paper award in KDD 2020	08/2020
CVPR 2023 Outstanding Reviewer Award	06/2023
NeurIPS 2021 Outstanding Reviewer Award	10/2021

FEATURED PUBLICATIONS

1. **Yang, H.,** Yin, H., Shen, M., Molchanov, P., Li, H., & Kautz, J. (2023). Global Vision Transformer Pruning with Hessian-Aware Saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18547-18557).
2. **Yang, H**.*, Liu, Y.*, Dong, Z., Keutzer, K., Du, L., & Zhang, S. (2023). NoisyQuant: Noisy Bias-Enhanced Post-Training Activation Quantization for Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20321-20330).
3. **Yang, H.***, Xiao, L.*, Dong, Z., Keutzer, K., Du, L., & Zhang, S. (2023). CSQ: Growing mixed-precision quantization scheme with bi-level continuous sparsification. In *2023 60th ACM/IEEE Design Automation Conference (DAC)* (pp. 1-6). IEEE.
4. **Yang, H.**, Yang, X., Gong, N. Z., & Chen, Y. (2022). HERO: Hessian-Enhanced Robust Optimization for Unifying and Improving Generalization and Quantization Performance. In *Proceedings of the 59th Annual Design Automation Conference* (pp. 25-30). (**Ranked first in the track**)
5. **Yang, H.**, Duan, L., & Li, H. (2021). BSQ: Exploring Bit-Level Sparsity for Mixed-Precision Neural Network Quantization. In *International Conference on Learning Representations*.
6. **Yang, H.**, Zhang, J., Dong, H., … & Li, H. (2020). DVERGE: Diversifying Vulnerabilities for Enhanced Robust Generation of Ensembles. In *Advances in Neural Information Processing Systems, 33,* 5505-5515. (**Oral**)
7. Li, A., Duan, Y., **Yang, H.**, Chen, Y., & Yang, J. (2020). TIPRDC: Task-Independent Privacy-Respecting Data Crowdsourcing Framework for Deep Learning with Anonymized Intermediate Representations. In *Proceedings of the 26th ACM SIGKDD* (pp. 824-832). (**Best student paper**)
8. **Yang, H.**, Tang, M., Wen, W., Yan, F., ... & Chen, Y. (2020). Learning Low-rank Deep Neural Networks via Singular Vector Orthogonality Regularization and Singular Value Sparsification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 678-679).
9. **Yang, H.**, Wen, W., & Li, H. (2020). DeepHoyer: Learning Sparser Neural Network with Differentiable Scale-Invariant Sparsity Measures. In *International Conference on Learning Representations*.

FULL PUBLICATIONS

Most up-to-date publication list can be found at https://scholar.google.com/citations?user=bjNCUt8AAAAJ

Conference and Workshop Proceedings

1. **Yang, H.***, Chen, A.*, Gan, Y., Gudovskiy, D., … & Keutzer, K. (2024). Split-Ensemble: Efficient OOD-aware Ensemble via Task and Model Splitting. In *International Conference on Machine Learning*. PMLR.
2. **Yang, H.,** Huang, Y., Dong, Z., … & Zhang, S. (2024). Fisher-aware Quantization for DETR Detectors with Critical-category Objectives. In *ICML'24 Workshop on Advancing Neural Network Training (WANT)*.
3. Huang, Q., **Yang, H.**, Zeng, E., & Chen, Y. (2024). A Deep-Learning-Based Multi-modal ECG and PCG Processing Framework for Label Efficient Heart Sound Segmentation. In *IEEE/ACM CHASE*.

4. Zhang, R., Luo, Y., Liu, J., **Yang, H.**, Dong, Z., … & Zhang, S. (2024). Efficient Deweahter Mixture-of-Experts with Uncertainty-Aware Feature-wise Linear Modulation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 15, pp. 16812-16820).
5. **Yang, H.,** Yin, H., Shen, M., Molchanov, P., Li, H., & Kautz, J. (2023). Global Vision Transformer Pruning with Hessian-Aware Saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18547-18557).
6. **Yang, H**.*, Liu, Y.*, Dong, Z., Keutzer, K., Du, L., & Zhang, S. (2023). NoisyQuant: Noisy Bias-Enhanced Post-Training Activation Quantization for Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20321-20330).
7. Li, X., Liu, Y., Lian, L., **Yang, H.**, Dong, Z., Kang, D., ... & Keutzer, K. (2023). Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 17535-17545).
8. Zhang, Y., Dong, Z., **Yang, H.**, Lu, M., Tseng, C. C., Du, Y., ... & Zhang, S. (2023). QD-BEV: Quantization-aware View-guided Distillation for Multi-view 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3825-3835).
9. **Yang, H.***, Xiao, L.*, Dong, Z., Keutzer, K., Du, L., & Zhang, S. (2023). CSQ: Growing mixed-precision quantization scheme with bi-level continuous sparsification. In *2023 60th ACM/IEEE Design Automation Conference (DAC)* (pp. 1-6). IEEE.
10. Yang, X., **Yang, H.**, Zhang, J., Li, H. H., & Chen, Y. (2022). On Building Efficient and Robust Neural Network Designs. In *2022 56th Asilomar Conference on Signals, Systems, and Computers* (pp. 317-321). IEEE.
11. **Yang, H.**, Yang, X., Gong, N. Z., & Chen, Y. (2022). HERO: Hessian-Enhanced Robust Optimization for Unifying and Improving Generalization and Quantization Performance. In *Proceedings of the 59th Annual Design Automation Conference* (pp. 25-30).
12. **Yang, H.**, Duan, L., & Li, H. (2021). BSQ: Exploring Bit-Level Sparsity for Mixed-Precision Neural Network Quantization. In *International Conference on Learning Representations*.
13. Chen, Y., Li, A., **Yang, H.**, Zhang, T., Yang, Y., Li, H., ... & Pajic, M. (2021). AI-Powered IoT System at the Edge. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)* (pp. 242-251). IEEE.
14. Yang, X., Belakaria, S., Joardar, B. K., **Yang, H.**, Doppa, J. R., Pande, P. P., ... & Li, H. H. (2021, November). Multi-objective optimization of ReRAM crossbars for robust DNN inferencing under stochastic noise. In *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)* (pp. 1-9). IEEE.
15. Xie, Z., Xu, X., Walker, M., Knebel, J., Palaniswamy, K., Hebert, N., Hu, J., **Yang, H.**, ... & Das, S. (2021, October). APOLLO: An automated power modeling framework for runtime power introspection in high-volume commercial microprocessors. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 1-14).
16. Zhang, J., **Huang, Y.**, Yang, H., Martinez, M., Hickman, G., Krolik, J., & Li, H. (2021, June). Efficient fpga implementation of a convolutional neural network for radar signal processing. In *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (pp. 1-4). IEEE.
17. Li, A., Guo, J., **Yang, H.**, Salim, F. D., & Chen, Y. (2021, May). Deepobfuscator: Obfuscating intermediate representations with privacy-preserving adversarial learning on smartphones. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation* (pp. 28-39).
18. Inkawhich, N., Liang, K. J., Zhang, J., **Yang, H.**, Li, H., & Chen, Y. (2021). Can Targeted Adversarial Examples Transfer When the Source and Target Models Have No Label Space Overlap?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 41-50).
19. Sun, J., Li, A., Wang, B., **Yang, H.**, Li, H., & Chen, Y. (2021). Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9311-9319).
20. **Yang, H.**, Zhang, J., Dong, H., … & Li, H. (2020). DVERGE: Diversifying Vulnerabilities for Enhanced Robust Generation of Ensembles. In *Advances in Neural Information Processing Systems, 33,* 5505-5515.
21. Li, A., Duan, Y., **Yang, H.**, Chen, Y., & Yang, J. (2020). TIPRDC: Task-Independent Privacy-Respecting Data Crowdsourcing Framework for Deep Learning with Anonymized Intermediate Representations. In *Proceedings of the 26th ACM SIGKDD* (pp. 824-832).

22. **Yang, H.**, Tang, M., Wen, W., Yan, F., ... & Chen, Y. (2020). Learning Low-rank Deep Neural Networks via Singular Vector Orthogonality Regularization and Singular Value Sparsification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 678-679).
23. **Yang, H.**, Wen, W., & Li, H. (2020). DeepHoyer: Learning Sparser Neural Network with Differentiable Scale-Invariant Sparsity Measures. In *International Conference on Learning Representations*.
24. Zhang, J., **Yang, H.**, Chen, F., Wang, Y., & Li, H. (2019). Exploring bit-slice sparsity in deep neural networks for efficient reram-based deployment. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)* (pp. 1-5). IEEE.
25. Cheng, H. P., Shen, J., **Yang, H.**, Wu, Q., Li, H., & Chen, Y. (2019). Adverquil: an efficient adversarial detection and alleviation technique for black-box neuromorphic computing systems. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference* (pp. 518-525).
26. Liu, X., **Yang, H.**, Liu, Z., Song, L., Li, H., & Chen, Y. (2019). Dpatch: An adversarial patch attack on object detectors. In *SafeAI 2019*.
27. Nixon, K. W., Mao, J., Shen, J., **Yang, H.**, Li, H. H., & Chen, Y. (2018). Spn dash-fast detection of adversarial attacks on mobile via sensor pattern noise fingerprinting. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (pp. 1-6). IEEE.
28. Song, C., Cheng, H. P., **Yang, H.**, Li, S., Wu, C., Wu, Q., ... & Li, H. (2018). MAT: A multi-strength adversarial training method to mitigate adversarial attacks. In *2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (pp. 476-481). IEEE.
29. Qiao, X., Cao, X., **Yang, H.**, Song, L., & Li, H. (2018). AtomLayer: A universal ReRAM-based CNN accelerator with atomic layer computation. In *Proceedings of the 55th Annual Design Automation Conference* (pp. 1-6).
30. Yuan, Z., Yue, J., **Yang, H.**, Wang, Z., Li, J., Yang, Y., ... & Liu, Y. (2018). Sticker: A 0.41-62.1 TOPS/W 8Bit neural network processor with multi-sparsity compatible convolution arrays and online tuning acceleration for fully connected layers. In *2018 IEEE symposium on VLSI circuits* (pp. 33-34). IEEE.

Journal Publications

1. Wu, X., Hanson, E., Wang, N., Zheng, Q., Yang, X., **Yang, H.**, … & Li, H. (2024). Block-Wise Mixed-Precision Quantization: Enabling High Efficiency for Practical ReRAM-based DNN Accelerators. *IEEE Transactions on Computer Aided Design of Integrated Circuits & Systems (TCAD)*
2. Yang, X., **Yang, H.**, Doppa, J. R., Pande, P. P., Chakrabarty, K., & Li, H. (2022). Essence: Exploiting structured stochastic gradient pruning for endurance-aware reram-based in-memory training systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.
3. Mao, J., **Yang, H.**, Li, A., Li, H., & Chen, Y. (2021). Tprune: Efficient transformer pruning for mobile devices. *ACM Transactions on Cyber-Physical Systems*, 5(3), 1-22.
4. Song, C., Cheng, H. P., **Yang, H.**, Li, S., Wu, C., Wu, Q., & Li, H. (2020). Adversarial attack: A new threat to smart devices and how to defend it. *IEEE Consumer Electronics Magazine*, 9(4), 49-55.

In Submission and Preprints

1. Liu, Y., Zhang, R., **Yang, H.**, Keutzer, K., Du, Y., Du, L., & Zhang, S. (2024). Intuition-aware Mixture-of-Rank-1-Experts for Parameter Efficient Finetuning. *arXiv preprint arXiv:2404.08985*.
2. Zhang, R., Cheng, A., Luo, Y., Dai, G., **Yang, H.**, Liu, J., ... & Zhang, S. (2024). Decomposing the Neurons: Activation Sparsity via Mixture of Experts for Continual Test Time Adaptation. *arXiv preprint arXiv:2405.16486*.
3. Ma, Z., Zhou, D., Yeh, C. H., Wang, X. S., Li, X., **Yang, H.**, ... & Feng, J. (2024). Magic-Me: Identity-Specific Video Customized Diffusion. *arXiv preprint arXiv:2402.09368*.
4. Zhang, R., Cai, Z., **Yang, H.**, Liu, Z., Gudovskiy, D., Okuno, T., ... & Zhang, S. (2024). VeCAF: VLM-empowered Collaborative Active Finetuning with Training Objective Awareness. *arXiv preprint arXiv:2401.07853*.
5. Zhang, J., **Yang, H.**, & Li, H. (2023). HCE: Improving performance and efficiency with heterogeneously compressed neural network ensemble. *arXiv preprint arXiv:2301.07794*.

Books and Book chapters

1. Li, A., **Yang, H.**, & Chen, Y. (2020). Task-Agnostic Privacy-Preserving Representation Learning via Federated Learning. In *Federated Learning* (pp. 51-65). Springer, Cham.

2. Chen, Y., Li, H., & **Yang, H.** (2023). Computer Engineering Machine Learning and Neural Networks (textbook for Duke ECE 661, in preparation)

SERVICES

Reviewer Service

- NeurIPS, ICLR, ICML, MLSys, CVPR, ICCV, AAAI, IJCAI, KDD, WACV
- IEEE TPAMI, IEEE TNNLS, IEEE TCASAI, TMLR, ACM TACO, ACM JETC, IEEE Access

Workshop Service

- Organizer, 3rd Workshop on Practical Deep Learning: Towards Efficient and Reliable LLMs @ IEEE CAI 2024
- Session Chair, 2nd International Workshop on Practical Deep Learning in the Wild @ AAAI 2023

Public Outreach

- Panelist for YICAI public interview on "Stories of the Generation Z AI Researchers" @ WAIC 2023
- Co-host and panelist for AI TIME public online seminar on "Towards Efficient DNN Architecture"

Educational Service

- Mentor in the 2023-2024 Berkeley AI Research (BAIR) undergrad mentorship program.
- Volunteered in the 2018 Females and Allies Excelling More in Math, Engineering, and Science (FEMMES+) Capstone event.

INVITED TALKS

Distribution-aware Post-training Quantization for Large Vision Language Models

- Invited talk at the 7th Workshop on Efficient Deep Learning for Computer Vision @ CVPR 2024, Seattle, WA

Exploring Bit-Level Patterns for Efficient NN Quantization and Deployment

- Invited talk at the 19th Embedded Vision Workshop @ CVPR 2023, Vancouver BC, Canada

Hero: Hessian-enhanced robust optimization for unifying and improving generalization and quantization

- Invited talk at ASP-DAC 2023 Designer's Forum, Tokyo, Japan (hybrid)

Robust DNN Inference under Input, Quantization and On-Chip Stochastic Noises

- Invited talk at CCF DAC 2021, Wuhan, China (hybrid)

DVERGE: diversifying vulnerabilities for enhanced robust generation of ensembles

- Invited online presentation at VALSE Student Webinar, AI TIME PhD Series, and TechBeat platforms.

RESEARCH FUNDINGS

Panasonic through BAIR Open Research Commons (2 projects per year) 06/2022 to 05/2024

- Controllable Quantization and Generalization for On-device Learning, $100,000
- Robust Neural Architecture Search with Improved Generalization on Corrupted Data, $100,000
- VLM-Empowered Collaborative Model Adaptation, $100,000
- Efficient LLM for multi-task specialization (tentative), $100,000

Defense Advanced Research Projects Agency (DARPA) Grant – HR00111990079 09/2019 to 03/2021

- Robust Ensemble Generation from Distilled Feature Transforms (REG-DFT), $299,579

Defense Advanced Research Projects Agency (DARPA) Grant – HR00112090054 04/2020 to 09/2021

- Neural-Network Enhanced Radar Surveillance (NNERS), $998,787

My research is further supported in part by the following grants, which my research outcomes contribute to.

National Science Foundation (NSF) Grant – 2112562 10/2021 to 05/2022

- AI Institute for Edge Computing Leveraging Next Generation Networks (Athena), $8,800,000

National Science Foundation (NSF) Grant – 1822085 09/2018 to 05/2022

- IUCRC for Alternative Sustainable and Intelligent Computing (ASIC), $7,500,000