Guihong Li

Email: lgh@utexas.edu

Education

- **The University of Texas at Austin** — Austin, Texas
  Ph.D. of Electrical and Computer Engineering; GPA: 4.0/4.0 — *August 2019 - Now*

  Advisor: *Radu Marculescu*

- **Tsinghua University** — Beijing, China
  Graduate student of Nano Integrated Circuits and Systems — *August 2018 - July 2019*

- **Beijing University of Posts and Telecommunications** — Beijing, China
  Bachelor's degree in Communication Engineering; GPA: 92.59/100; Rank: 6/565 — *September 2014 - June 2018*

RESEARCH EXPERIENCE

My research focuses on developing **efficient** and **trustworthy** deep neural networks:

- Enhancing the trustworthiness of machine learning by exerting control over the training data and generated content in Generative AI.
- Improving the time efficiency of AutoML through the investigation of multiple explainability aspects, particularly in the development of theoretically-grounded, training-free NAS approaches.
- Facilitating automatic dynamic neural network design, taking into account hardware resource availability.
- Enabling real-time inference and training on budget-friendly edge devices through network-system co-design.

Industry Experience

- **JPMorgan Chase & Co** — New York
  Research Intern (Full-time) — *June 2023 - October 2023*
  Supervisor: Dr. Richard CF. Chen, Dr. Hsiang Hsu
  - **Trustworthy Generative models**: Control the contents generated by image generative models.
  - **Efficient Machine Unlearning**: Build a efficient machine unlearning algorithm to quickly remove the information from a trained model.

- **ARM ML Tech** — San Jose
  Research Intern (Full-time) — *May 2021 - August 2021*
  Supervisor: Dr. Kartikeya Bhardwaj, Dr. Naveen Suda, Dr. Lingchuan Meng
  - **Hardware Performance evaluation**: Build a model to estimate neural networks' latency on neural accelerators.
  - **Hardware-aware NAS**: Explore the neural architecture search technique to search for hardware-efficient models.

Selected Publications

- Guihong Li, Hsiang Hsu, Chun-Fu Chen, Radu Marculescu."Machine Unlearning for Image-to-Image Generative Models." submitted to ICLR 2024.
- Guihong Li, Duc Hoang, Kartikeya Bhardwaj, Ming Lin, Zhangyang Wang, Radu Marculescu. "Zero-Shot Neural Architecture Search: Challenges, Solutions, and Opportunities." submitted to IEEE T-PAMI.
- Guihong Li, Kartikeya Bhardwaj, Yuedong Yang, and Radu Marculescu. "TIPS: Topologically Important Path Sampling for Anytime Inference Networks." ICML 2023.
- Guihong Li, Yuedong Yang, Kartikeya Bhardwaj, and Radu Marculescu. "ZiCo: Zero-shot NAS via inverse Coefficient of Variation on Gradients." ICLR 2023 (**Spotlight**).
- Guihong Li, Sumit K. Mandal, Umit Y. Ogras, and Radu Marculescu. "FLASH: Fast Neural Architecture Search with Hardware Optimization." CODES+ISSS 2021.
- Kartikeya Bhardwaj*, Guihong Li*, and Radu Marculescu. "How does topology influence gradient propagation and model performance of deep networks with DenseNet-type skip connections?" CVPR 2021. (*Co first author)
- Dawei Liang*, Guihong Li*, Rebecca Adaimi, Radu Marculescu and Edison Thomaz. "AudioIMU: Enhancing Inertial Sensing-Based Activity Recognition with Acoustic Models." ISWC 2022. (*Co first author) **Best paper nomination**
- Yuedong Yang, Guihong Li, and Radu Marculescu. "Efficient On-device Training via Gradient Filtering." CVPR 2023.
- A. Alper Goksoy, Guihong Li, Sumit K. Mandal, Umit Y. Ogras, Radu Marculescu. "CANNON: Communication-Aware Sparse Neural Network Optimization." IEEE TETC 2023.
- Yuedong Yang, Hung-Yueh Chiang, Guihong Li, Diana Marculescu, Radu Marculescu. "Efficient Low-rank Backpropagation for Vision Transformer Adaptation." NeurIPS 2023.

Selected Projects

- **Contents Control for Generative AI:** Build an end-to-end framework to avoid generating unwanted contents by generative models, consisting of: (*i*) theoretically analyze the problem and derive and unique and optimal solution to this problem; (*ii*) design an novel algorithm to efficiently implement the obtained theoretical solution; (*iii*) comprehensive evaluation on large-scale experimental setup and for all mainstream generative models.

- **Zero-shot NAS framework:** Developed an end-to-end hardware-aware NAS framework for mobile devices, consisting of: (*i*) converting PyTorch models to TF-lite models and deploying them on ARM-based devices; (*ii*) profiling model execution, constructing models to estimate hardware performance; (*iii*) integrating hardware performance models into the search space, and conducting searches using the proposed proxy; (*iv*) training and deploying the searched model on the target device.

- **Structural pruning with computation dependency:** Designed an end-to-end structural pruning method considering layer-wise computation dependency, involving: (*i*) pruning the network structurally and fine-tuning the pruned network; (*ii*) extracting computation dependency by building layer-wise computation graphs; (*iii*) further pruning unuseful channels based on the computation graph.

- **CNN Compilation for PIM:** Created an end-to-end compiler for CNN inference on Processing-in-memory (PIM) hardware, comprising: (*i*) constructing layer-wise computation graphs for given CNNs; (*ii*) partitioning each layer's computation operator into multiple arrays (kernels) based on PIM array size; (*iii*) generating execution code by combining each layer's compilation and skip connections.

- **Quantization with variable bit widths:** Developed a quantization-aware training algorithm, featuring: (*i*) introducing rescaling techniques during backward and forward propagation by analyzing training dynamics to facilitate convergence; (*ii*) proposing a channel-level bit-width-variable quantization scheme and implementing rescaling techniques based on PyTorch.

- **On-device model personalization:** Designed a customized backward propagation method for efficient on-device model personalization, consisting of: (*i*) proposing memory and computation-efficient backward propagation by introducing approximation during training; (*ii*) implementing the framework by modifying the operator of backward computation based on CUDNN and MKLDNN, and verifying hardware efficiency.

- **Efficient & Robust Visual Wake Words:** Created an efficient and robust model for Visual Wake Words, including: (*i*) developing an automatic tool to collect diverse samples from the internet under various scenarios and labeling them without human intervention; (*ii*) training the model and conducting hard negative and hard positive mining to enhance robustness.

- **Human Activity Recognition:** Devised a novel approach to augment inertial measurement unit (IMU) models for human activity recognition (HAR) with superior acoustic knowledge of activities, involving: (*i*) proposing a teacher-student framework to derive an IMU-based HAR model, incorporating an advanced audio-based teacher model to guide the student HAR model; (*ii*) deploying the HAR model for inference on wearable devices equipped with IMU and microphones, using only motion data as input.

Honors and Awards

- Premier Scholarship Candidate (Highest college honor. I am the only junior student; the rest are all seniors) - 2016
- National Scholarship (Top 1%) - 2016, 2017
- National Encouragement Scholarship (Top 2%) - 2015
- Ranked 1st in the nationwide final of China Next-Generation Network Technology Innovation Contest - 2015
- First prize (Top 10%) among almost 10,000 teams at The international Mathematics Competition in modeling

Skills Summary

• **Language**	Python, C/C++, Matlab
• **Frameworks**	TensorFlow, PyTorch, Scikit-learn, Scipy, TVM, ONNX, Keil Studio