Ruijie Zhang

📱 +86 186-7157-3524 | @ zrjhust@gmail.com | GitHub

Research Interests

My research interests focus on efficient and responsible machine learning, AI for science, computer vision, and the mathematical & physical principles behind them. Specifically, I am interested in model compression methods, including network pruning, distillation, and feature reduction; On-device training and inference; Machine learning for pathology, etc. I have recently focused on

- LLMs/LVMs Efficient Pretrain
- Model Compression: few-shot pruning methods, pruning ratio theoretical limitation, feature reduction for model-acceleration.
- AI for Science: accurate & responsible segmentation and classification models for medical images
- Computer Vision: image understanding, including gigapixel images classification & segmentation & reconstruction

Education

Huazhong University of Science and Technology Wuhan, China
M.S. in School of Electronic Information and Communication; *Sep 2021 – Sep 2024 (Expected)*
B.S. in School of Electronic Information and Communication; *Sep 2015 – Jun 2019*

Hiroshima University Hiroshima, Japan
Exchange Student in School of Engineering *Sep 2018 – Jun 2019*

University of California, Santa Barbara
Ph.D. in Dept. of Computer Science, Supervised by Prof. Zheng Zhang *Sep 2024 –*

Research Experience

How Sparse Can We Prune A Deep Network: A Fundamental Limit Perspective GitHub
Supervised by Prof. Jun Sun *Jun 2021 – May 2023*

- By merely enforcing the sparsity constraint on the original loss function, we're able to characterize the sharp phase transition point of the pruning ratio, which corresponds to the boundary between the feasible and the infeasible, from the perspective of high-dimensional geometry.
- We found the phase transition point of the pruning ratio equals the squared Gaussian width of some convex body, normalized by the original dimension of parameters.
- As a byproduct, we provide a novel network pruning algorithm which is essentially a global one-shot pruning one.
- We provide efficient countermeasures to address the challenges in computing the involved Gaussian width, including the spectrum estimation of a large-scale Hessian matrix and dealing with the non-definite positiveness of a Hessian matrix.

Multi-level Multiple Instance Learning with Transformer GitHub
Supervised by Prof. Xinggang Wang *Oct 2022 – May 2023*

- A Multi-level MIL with Transformer (MMIL-Transformer) approach is proposed. By introducing a hierarchical structure to MIL, this approach enables efficient handling of MIL tasks that involve a large number of instances.
- Conducted a set of experiments on Whole slide image classification task(CAMELYON16 Dataset and TCGA Dataset), where MMIL-Transformer demonstrate superior performance compared to existing SOTA methods.

Compressive Hyperspectral Image Reconstruction. GitHub
Supervised by Prof. Xinggang Wang *Aug 2022 – Apr 2023*

- Based on range-null space decomposition, a novel framework named RND-SCI for compressive hyperspectral image (HSI) reconstruction is proposed. RND-SCI significantly boosts the performance of HSI reconstruction networks in retraining, fine-tuning or plugging into a pre-trained off-the-shelf model.
- Incorporate SAUNet, an extremely fast HSI reconstruction network is designed, RND-SAUNet, which achieves 91 frames per second while maintaining superior reconstruction accuracy compared to other less time-consuming methods.

Projects

Deep Learning and Tele-Robotic

- Participated in a workshop at the National University of Singapore and received an A grade for designing an automated replenishment robot.
- Based on TCP/IP, this robot can communicate with a cloud host through a video module mounted on a Raspberry Pi. The cloud server processes the video signals using MASK-RCNN to identify different quantities of various products. If the product quantity drops below a threshold, the robot autonomously tracks and avoids obstacles using a radar module in conjunction with the video module, ultimately completing the replenishment task.

Design of CPU based on MIPS

- Utilizing HDL and the Xilinx FPGA platform, the task involved crafting a single-cycle MIPS processor designed to execute the specified MIPS instruction set.
- The processor supports basic arithmetic and logic operations such as add, sub, and, or, slt; basic memory operations like lw, sw.; basic program control with beq and j.

Super-Resolution Localization in Mobile Scenarios

- As the team leader, I was responsible for proposing models and algorithms that utilize FFT and MUSIC to detect the distance and angle of objects. In order to simultaneously address the requirements of high precision and low complexity, we introduced the high-precision MUSIC algorithm based on fuzzy matching.
- We provided theoretical justification for the limited resolution of the MUSIC algorithm in certain scenarios (high-dimensional environments where the number of antennas and samples grows proportionally) and made adjustments based on random matrix theory to enhance algorithm performance.
- We developed a method based on the Transformer architecture. However, the model's performance was constrained due to the limited amount of available data. To address this limitation, we employed a diffusion model to augment the existing dataset, leading to a substantial enhancement in the model's performance.

Awards & Achievements

- **Second Price of 19th China Post-Graduate Mathematical Contest in Modeling**
- **Japan JASSO Scholarship(2018)**
- **Academic Scholarship(2021,2022)**
- **Outstanding Graduates Award (Undergraduate).**

Publications

Zhang Q, Zhang R, Sun J, et al. How Sparse Can We Prune A Deep Network: A Geometric Viewpoint[J]. arXiv preprint arXiv:2306.05857, 2023.(submitted to ICML 2024)

Zhang R, Zhang Q, Liu Y, et al. Multi-level multiple instance learning with transformer for whole slide image classification[J]. arXiv preprint arXiv:2306.05029, 2023(submitted to IEEE Transactions on Medical Imaging (TMI)).

Wang J, Wang S, Zhang R, et al. A Range-Null Space Decomposition Approach for Fast and Flexible Spectral Compressive Imaging[J]. arXiv preprint arXiv:2305.09746, 2023.(submitted to ICCV 2023)

Skills

Programming: Python, Pytorch, C, C++, Assembly, Verilog, HTML

Technologies: Git, Arduino, Xilinx ISE

Languages: Chinese (Native), English, Japanese

More About Me

"There is only one heroism in the world: to see the world as it is, and to love it." — Romain Rolland

Besides my journey to science, I am also on the journey to myself.

I am a solitary adventurer on two wheels. I cycled across the vast expanse of China, traversing the length of National Highway 318 from Shanghai to Tibet.

I am a part-time photographer. In addition to capturing the grandeur of mountains, the serenity of rivers, the enchantment of skies, and the allure of oceans, I managed to partly support my college education and daily expenses through my photography work.

I am a ukulele player. The strings of this small instrument have been my companions on countless nights, strumming away my worries and evoking a sense of calm and joy.

I have long held the belief that the essence of life is found in the act of experiencing, exploring the world, immersing oneself deeply in its wonders, and loving it wholeheartedly.