Shouyi Li

sl5632@columbia.edu | +1 (323)-633-8029 | linkedin.com/in/shouyili

EDUCATION

Columbia University — New York, NY
- MS in Computer Science (Machine Learning Track) — Aug 2024 – Dec 2025
- Courses: Machine Learning Theory, Natural Language Processing, Computer Vision, Cloud Computing

University of Southern California — Los Angeles, CA
- BS in Computer Engineering and Computer Science, GPA: 3.9/4.0 — Aug 2020 – May 2024
- Teaching Assistant (TA): Artificial Intelligence, Digital Circuits, Discrete Math

WORK EXPERIENCE

Meituan — Beijing, China
LLM Inference Architecture Engineer Intern — May 2024 – Aug 2024
- Conducted batch inference optimizations for large-scale LLM deployment, focusing on pruning techniques
- Implemented weight and token pruning with customized CUDA compute kernels, reducing MLP memory usage and latency by 26% during inference; initiated team-wide discussions to communicate deployment team downstream
- Integrated research prototypes into production applications, leading processor and system performance modeling
- Read latest research on optimized inference architecture and shared insights with team, gaining expertise in advanced techniques, including FlashAttention, PagedAttention, speculative decoding, MoE, MLA, RAG, and more

Hongmeng Measurement and Control Technology Co., Ltd — Xuzhou, China
Software Engineering Intern — Feb 2022 – Aug 2022
- Functioned as a software developer and tester for the Furnace Safety Supervision System APP; collaborated with senior engineers to improve system performance and interacted with hardware teams to resolve integration issues
- Applied the Modbus protocol and CRC-16 algorithm for data transmission, guaranteed 100% error bit detection, proposed a lookup table approach to speed up CRC algorithm's computation by 530% on average

RESEARCH EXPERIENCE

University of Southern California, FPGA/High Performance Computing Lab — Los Angeles, CA
Student Researcher — Mar 2023 – Aug 2024
- Engaged in project "*Accelerating ViT Inference on FPGA through Static and Dynamic Pruning*" (FCCM 2024)
- Incorporated block pruning and token dropping techniques to accelerate Vision Transformer model inference
- Designed task-specific FPGA accelerator to address inherent computational challenges in ViTs
- Reduced computation complexity by 3.4× with ≈ 3% accuracy drop and a model compression ratio of 1.6×
- Attained 12.8×, 3.2×, 2.1× speedup compared with state-of-the-art implementation on CPU, GPU, and FPGA

University of Southern California, Convergent Science Institute in Cancer — Los Angeles, CA
Student Researcher — Aug 2023 – Aug 2024
- Developed project "*Deep Morphological Profiling of Immune Cells in Peripheral Blood*" under mentor's guidance
- Built a Convolutional Neural Network (CNN) model to deconvolve immune repertoire from IF image data and predict immune cell subclasses to up to 5 types of antibodies, achieved up to 91% accuracy
- Utilized unsupervised learning to identify 6 clusters of immune cells based on latent feature space
- Introduced representation learning to characterize morphological features of distinct immune cells

PROJECTS

CampusHub — Aug 2024 – Dec 2024
- Design a microservices-based school event app, utilizing React, FastAPI, and MySQL, deployed on AWS
- Lead integration of RDS database using AWS DBaaS, ensuring robust data management and scalability
- Employ deployment across multiple environments, including VMs, containers, and cloud services

OpenHome — Jan 2024 – June 2024
- Contributed to an intelligent AI speaker design capable of performing more than 100 tasks based on commands
- Pruned and quantized LLM to increase speed by 40%, leading team discussions on performance improvements

Planner Pal — Sept 2022 – Dec 2022
- Developed a multifunctional web app extension; helped over 200 users stay organized with day-to-day tasks
- Built backend APIs using Spring-boot, with JS as frontend, Google Firebase as database, and led deployment on Google Compute Engine; engaged with users to gather feedback and enhance app's features

SKILLS

Languages: C++/C, Python, Java, HTML&CSS, Javascript, SQL, Verilog, Assembly
Frameworks/Libraries: PyTorch, TensorFlow, CUDA, Spring-boot, REST API, FastAPI, React, Flask, Triton, HPC
Cloud/Databases: AWS, GCP, Microsoft Azure, MySQL, Firebase, MongoDB
Other Tools: Linux, Git, Docker, Bash/Shell, VM, VS Code, PyCharm, Android Studio, IntelliJ, DataGrip, Jupyter