Susav Shrestha

Santa Clara, California | (682) 227-7829 | susavlsh10@gmail.com | LinkedIn | GitHub | susavlsh10.github.io

AI systems researcher with hands-on experience in scalable LLM inference, sparse computing, and custom GPU kernel development. Passionate about bridging ML algorithms with systems for high-throughput deployments.

EDUCATION

Texas A&M University, College Station, Texas — Est. Dec 2025
PhD in Computer Engineering, GPA 4.0

University of Texas Arlington, Arlington, Texas — Aug 2017 - May 2021
Bachelors in Electrical Engineering with Honors, Minor in Computer Science, GPA 3.97
Awards: Cambridge Learners Award 2015, Academic Excellence Award 2016, Innovation Award 2021

WORK EXPERIENCE

Texas A&M University, *Graduate Research Assistant*, College Station, TX — Aug 2021 - Present
- Led efforts to design efficient systems for deep learning applications; Neural Information Retrieval (IR), LLMs.
- Built SSD-based IR system with custom GPU kernels and high-hit-rate prefetcher (16x memory reduction)
- Accelerated LLM inference using adaptive speculative decoding and contextual sparsity.

NVIDIA, *Research Intern*, Santa Clara, CA — May 2025 - Present
- Designed Hybrid Attention MLP Parallelism (HAMP) to accelerate LLM inference in disaggregated GPU nodes.
- Optimized expert dispatch and combine to scale MoE inference across GPU nodes efficiently.

NVIDIA, *Research Intern*, Austin, TX — May 2024 - Aug 2024
- Led research to accelerate LLM inferencing via activation and contextual sparsity in OPT, Llama 2 & 3 models.
- Developed custom selective GPU kernels to accelerate MLP layers by 1.5-3x and Attention layers up to 2.5x.
- Delivered **2.2x** end-to-end decoding speedup across various batch sizes and sequence lengths.

Samsung Semiconductor, *Research Intern*, San Jose, CA — May 2022 – Aug 2022
- Designed a SmartSSD-based Neural IR system with custom FPGA similarity kernels. (2 patent applications)
- Reduced CPU workload by 4x, reduced memory usage by 82% for IR using SSD offloading, near storage processing.

SELECTED PUBLICATIONS

- **Shrestha, S.** "*ESPN*: Memory Efficient Multi-Vector Information Retrieval*"* ISMM, 2024.
- **Shrestha, S.** "*Polar Sparsity:* High Throughput Batched LLM Inferencing with Scalable Contextual Sparsity" [Code]

RELEVANT PROJECTS

ML, Computer Vision, NLP, RL — Aug 2022 - Dec 2024
- Remodeled and trained the ResNet CNN model with Squeeze/Expand layers with accuracy of 88-92% in CIFAR-10.
- Built a music recommendation system using BM25, Content-based, Collaborative filtering & Matrix Factorization.
- Fine-tuned BERT for sentiment analysis, LLMs like GPT-2, LLaMA (LoRA) for text classification as text generation.
- Fine-tuned (LoRA) Stable Diffusion with multi objective reinforcement learning for enhanced image generation.

Tiny Social Network, Distributed Processing Systems [Code] — Dec 2023
- Implemented scalable, fault-tolerant distributed service in C++ using gRPC with follow and timeline features.

TECHNICAL SKILLS

- *Technical:* C/C++, Python, CUDA, Triton, MATLAB, OpenCL, OpenMP, Pthreads, MPI, Pytorch, Scikit-learn, gRPC, Hadoop, LangChain, Azure, AWS, Verilog, VHDL, Vitis/Vivado HLS, Object Oriented (OOP), GIT, Linux.