

Susav Shrestha

College Station, Texas | (682) 227-7829 | susavsh10@gmail.com | [LinkedIn](#) | [GitHub](#) | susavsh10.github.io

EDUCATION

Texas A&M University, College Station, Texas
PhD in Computer Engineering, GPA 4.0

Est. May 2026

University of Texas Arlington, Arlington, Texas
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

Aug 2017 - May 2021

WORK EXPERIENCE

Texas A&M University, *Graduate Research Assistant*, College Station, TX

Aug 2021 - Present

- Led efforts to create more efficient systems for deep learning applications; Neural Information Retrieval (IR).
- Developed a novel retrieval architecture with Nvidia GDS, custom CUDA kernels and embedding prefetcher (90%+ hit rate) achieving 23% speedup and 16x lower memory footprint.
- Accelerated language models using adaptive speculative decoding

NVIDIA, *Research Intern*, Austin, TX

May 2024 - Aug 2024

- Led research to accelerate LLM inferencing via activation and contextual sparsity.
- Built sparsely activated OPT and Llama models by training activation routers for MLP and Attention layers.
- Developed custom sparse GPU kernels to accelerate MLP layers by 1.5-3x and Attention layers up to 2.5x.
- Delivered end-to-end decoding speedup of 1.6-2x across various batch sizes and sequence lengths.

Samsung Semiconductor, *Research Intern*, San Jose, CA

May 2022 – Aug 2022

- Developed a novel IR embedding retrieval and processing architecture, leveraging SmartSSD with custom similarity computation kernels (FPGA) to offload similarity processing close to storage.
- Reduced CPU workload by 4x, reduced memory usage by 82% without degrading retrieval quality of the IR model.
- Delivered 2 patents applications for efficient Neural Information Retrieval system design during the internship.

SELECTED PUBLICATIONS

- **Shrestha, S.**, Annapareddy, N., Li, Z. "ESPN: Memory Efficient Multi-Vector Information Retrieval." ISMM, 2024.
- **Shrestha, S.**, Settlemeyer, B., Dryden, N., & Annapareddy, N. "PolarInfer: High Throughput Batched LLM Inferencing with Scalable Contextual Sparsity." (Under preparation, 2025)

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

Dec 2023

- Developed a scalable, fault tolerant (multiple servers, coordinators) distributed service with follow, unfollow, list, timeline features built on top of gRPC in C++.

TECHNICAL SKILLS

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