Susav Shrestha

• LinkedIn • Website • sls7161@tamu.edu, susavlsh10@gmail.com • GitHub • 682-227-7829

EDUCATION

Texas A&M University **College Station, TX**
***Doctor of Philosophy* in Computer Engineering**. GPA 4.0 **Est. May 2025**
***Thesis*:** Hardware Efficient ML System Design

University of Texas Arlington **Arlington, TX**
***Bachelor of Science* in Electrical Engineering** with Honors, Minor in Computer Science. GPA 3.97 **May 2021**
Awards: Cambridge Learners Award 2015, Academic Excellence Award 2016, Dean's List 2018-21, Innovation Award 2021

PROFESSIONAL EXPERIENCE

Texas A&M University **College Station, TX**
PhD, Graduate Researcher **Aug 2021– Present**

- Led efforts to create more efficient systems for deep learning applications; Neural Information Retrieval (IR).
- Trained a BERT based model for IR using pytorch and knowledge distillation with Recall@1K of 96%.
- Developed a novel embedding processing system with Nvidia GDS & custom CUDA kernels achieving 23% speedup.
- Designed a flexible software prefetcher for Neural IR systems achieving hit rates exceeding 90%, allowing embeddings to be retrieved from storage with near memory latency with 16x less memory footprint.

Samsung Semiconductor Inc. **San Jose, CA**
Machine Learning Systems Research Intern **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.
- Accelerated neural inference by 64% by compressing the model with knowledge distillation, pruning and quantization.
- Delivered 2 patents applications for efficient Neural Information Retrieval system design during the internship.

University of Texas Arlington **Arlington, TX**
Research and Teaching Assistant **May 2019– Aug 2021**

- Developed signal processing algorithms to measure dynamic soil properties using a Radar system.
- Teaching assistant for course EE2347; instructor for Lab: Introduction to python and algorithms; graded assignments.

PUBLICATIONS

Shrestha, S., Annapareddy, N., & Li, Z. (2024). ESPN: Memory Efficient Multi-Vector Information Retrieval. (submission under review)

RELEVANT PROJECTS

Machine Learning and Computer Vision (CV) (Python)

- Remodeled and trained the ResNet CNN model with Squeeze/Expand layers with accuracy of 88-92% in CIFAR-10.
- Developed ML algorithms like Naïve Bayes, Decision trees & forests, Linear and Logistic Regression, Neural Networks, K-means clustering from scratch (no libraries) with > 90% accuracy.

Recommendation systems and Natural Language Processing (NLP)

- Built a music recommendation system using BM25, Content-based, Collaborative filtering & Matrix Factorization.
- Fine-tuned LLMs (Full, LoRA) like BERT & GPT-2 models for sentiment analysis, text classification as generation.

Distributed and Parallel Processing Systems (C/C++)

- Implemented Strassen's recursive matrix multiplication algorithm in CUDA and obtained 172x speedup over CPU.
- **Tiny Social Network**: Developed a scalable, fault tolerant (multiple servers, coordinators) distributed service with follow, unfollow, list, timeline features built on top of gRPC in C++.

SKILLS AND COURSES

Relevant Coursework: Advance Computer Architecture, Parallel Computing, Distributed Processing, Deep Learning, Machine Learning, NLP, Information Retrieval, Memory & Storage Systems, Operating Systems, Advanced Algorithms.
Technical: C/C++, Python, Java, CUDA, MATLAB, OpenCL, OpenMP, Pthreads, MPI, Pytorch, Scikit-learn, gRPC, Hadoop, Azure, AWS, Spark, Verilog, VHDL, Vitis/Vivado HLS, Object Oriented Programming (OOP), GIT, Linux.