Neelesh Gupta

neelesh-gupta23 | neeleshg23.github.io | neeleshg@usc.edu | +1 (832) 591 8299

Education

University of Southern California — **Los Angeles, CA**
M.S. Computer Science — *August 2024 - December 2025 (Expected)*

University of Southern California — **Los Angeles, CA**
B.S. Computer Science, Minor: Mathematics, GPA: 3.62/4 — *August 2021 - May 2025*

Work Experience

Research Assistant — **Los Angeles, CA**
USC Ming Hsieh Department of ECE - Parallel Computing/Data Science Lab — *January 2023 - Present*

- Develop novel techniques for accelerating deep learning models on specialized hardware accelerators for compute-intensive applications under the supervision of Prof. Viktor Prasanna and Prof. Raj Kannan.
- Mentor exchange students on research projects related to high-performance computing and hardware acceleration, providing guidance on problem formulation, algorithm design, and experimental setup.
- Organize and actively participate in weekly group meetings to discuss progress and exchange ideas.
- Present findings at conferences to showcase the lab's work and engage with the broader community.

Undergraduate Research Assistant — **Marina del Rey, CA**
USC Information Sciences Institute - STEEL: Security Research Lab — *May 2022 - January 2023*

- Developed scripts to automate the PCA algorithm for detecting energy grid outages in synthetic data.
- Performed sentiment analysis on the Enron corpus to train a machine learning model to classify spam.

Projects

Machine Learning for Data Prefetching — *Jan. 2023 – Present*

- Develop novel computer architecture approaches to tackle post-Moore's era memory subsystems overhead using machine learning to better cache utilization compared to traditional rule-based prefetchers.
- Compress neural network-based memory access prediction models using knowledge distillation, optimizing them for irregular memory access patterns prevalent in sparse data and graph analytics.
- Implement the first instance of a practical neural network-based prefetcher with inference latency comparable to rule-based prefetchers by distilling and approximating Transformers using tables.

Efficient Table-Based Neural Network — *Jan. 2023 - Present*

- Designed primitives for matrix multiplication, convolutions, and multi-headed self-attention that are approximated using a product quantization strategy and can be efficiently looked up in a table.
- Reduced arithmetic operations during inference by over 90% for table-lookup approximated operations.
- Developed a comprehensive design methodology and fine-tuning process for training full neural networks based on table-based primitives, using layer-masking and layer-wise table size hyperparameter tuning.

Skills

Programming Languages:	Python, C++, Java, Go, JavaScript
Parallel Computing & ML:	CUDA, OpenMP, MPI, PyTorch, Scikit-learn, NumPy
FPGA Development:	Vitis, Vivado, Pynq, Verilog, VHDL
Other Tools:	Linux, Bash, Git, LaTeX, Jupyter Notebooks, Matplotlib, Seaborn

Publications

[1] Neelesh Gupta, Narayanan Kannan, Pengmiao Zhang, and Viktor Prasanna. "TabConv: Low-Computation CNN Inference via Table Lookups". In: *Proceedings of the 21st ACM International Conference on Computing Frontiers*. CF '24. Ischia, Italy: Association for Computing Machinery, 2024. ISBN: 9798400705977. DOI: 10.1145/3649153.3649212. URL: https://doi.org/10.1145/3649153.3649212.

[2] Pengmiao Zhang, Neelesh Gupta, Rajgopal Kannan, and Viktor K. Prasanna. "Attention, Distillation, and Tabularization: Towards Practical Neural Network-Based Prefetching". In: *2024 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. San Francisco, CA, USA, May 2024, pp. 1–10. arXiv: 2401.06362 [cs.NE].

[3] Neelesh Gupta, Pengmiao Zhang, Rajgopal Kannan, and Viktor K. Prasanna. "PaCKD: Pattern-Clustered Knowledge Distillation for Compressing Memory Access Prediction Models". In: *IEEE High Performance Efficient Computing (HPEC)*. IEEE. Boston, MA, Sept. 2023.

[4] Jeffrey Liu, Rajat Tandon, Uma Durairaj, Jiani Guo, Spencer Zahabizadeh, Sanjana Ilango, Jeremy Tang, Neelesh Gupta, Zoe Zhou, and Jelena Mirkovic. "Did your child get disturbed by an inappropriate advertisement on YouTube?" In: *Proceedings of KDD Undergraduate Consortium*. KDD-UC'22. ACM. Washington, D.C., Oct. 2022.