Ahan Gupta

Email : ahangupta.96@gmail.com
Linkedin: https://www.linkedin.com/in/ahan-gupta-405619103/
Mobile : +1-415-966-5501
Github: https://github.com/spikerheado1234

Education

- **University of Illinois Urbana-Champaign** — Champaign, IL
  PhD in Computer Science — *Aug 2022 - Present*
- **National University of Singapore** — Singapore
  Bachelor of Computing in Computer Science — *Aug 2017 - Dec 2021*

Research Statement

I am broadly interested in researching high-performance Compiler & System level abstractions to accelerate deep-learning applications. My work melds both theory and practice, providing high-performance abstractions and systems that have strong theoretical guarantees.

Experience

- **Google DeepMind** — Mountain View, CA
  Student Researcher — *May 2024 - Present*
  - Investigated KV-cache compression strategies at scale.
  - Developed novel KV-cache compression algorithm that compresses KV-cache across the sequential dimension.
  - Debugged, trained and evaluated models at the billion parameter scale, assessing the strengths and weaknesses of posited strategy.

- **Citadel** — Hong Kong
  Software Engineering Intern — *May 2021 - Aug 2021*
  - Designed an authentication library to enable developers to integrate authentication logic with services in an easy manner
  - Contributed to a tool that monitors AWS usage of different desks
  - Built an extensible, general purpose, monitoring tool that enables developers and traders to track a wide range of internal services' uptime and accuracy

- **Google** — Singapore
  Software Engineering Intern — *May 2020 - Aug 2020*
  - Designed Asynchronous Web APIs via OpenAPI for authorisation microservice
  - Designed database Schemas for an authorisation microservice
  - Built Infrastructural groundwork to enable integration testing with databases
  - Implemented APIs that enable secure FIDO signature validation in HapiJS and TypeScript
  - Merged all code into production; Strategized to make all third party payment applications (e.g. Venmo, UPI) interoperable to the larger MojaLoop network in a secure manner, in subsequent builds

Publications

- *Ahan Gupta*, Yueming Yuan, Devansh Jain, Yuhao Ge, David Aponte, Yanqi Zhou, Charith Mendis (2024). SPLAT: Optimized GPU code generation framework for SParse reguLar ATtention. (In Submission). Preprint link: https://arxiv.org/abs/2407.16847
- *Ahan Gupta*, Yueming Yuan, Yanqi Zhou, Charith Mendis (2024). FLuRKA: Fast fused Low-Rank & Kernel Attention. (In Submission). Preprint link: https://arxiv.org/abs/2306.15799

Skills Summary

- **Languages**: Java, C++, Python, C, SQL, Javascript, Scala, Cuda
- **Tools**: Docker, Pytorch, Tensorflow, JAX, Keras, LLVM