Huy Nguyen

Eugene, Oregon huy.nguyenhuu.hust@gmail.com (458) 245-7465 My Website My LinkedIn
HuyNguyen-hust

Research Summary

My background in Natural Language Processing (NLP) encompasses Information Extraction (IE) and Continual Learning, which provided a foundational understanding of language models. I've since transitioned my focus to the cutting-edge domain of efficient Large Language Models (LLMs). My current research is dedicated to accelerating LLMs through low-level performance optimization, where I leverage a strong command of CUDA and Triton. I'm particularly interested in manually scheduling work at the instruction level to maximize GPU utilization and minimize idle resources during both training and inference, for both standard and quantized models.

Education

University of Oregon *Sept 2024 – Present*
PhD in Computer Science

- GPA: 4.08/4.0, Excellent Degree
- Research Topic: LLMs for Low-Resource Languages, Efficient LLMs
- Supervisor: Assoc. Prof. Thien Huu Nguyen, thienn@uoregon.edu

Hanoi University of Science and Technology *Sept 2018 – May 2023*
BS in Computer Science

- GPA: 3.71/4.0, Excellent Degree
- Advisor: Dr. Linh Ngo Van, linhnv@soict.hust.edu.vn

Experience

Microsoft *Research Intern* *Redmond, WA, USA Jun 2025 – Present*

- Project: Developed and optimized scheduling methods on NVIDIA GPUs for fast inference in Large Language Models (LLMs).

—

VinAI Research *AI Resident* *Hanoi, Vietnam Jul 2022 – Jul 2024*

- Supervisors:
  - Asst. Prof. Luu Anh Tuan, anhtuan.luu@ntu.edu.sg
  - Assoc. Prof. Thien Huu Nguyen, thienn@uoregon.edu
- Projects:
  - **A Spectral Viewpoint on Continual Relation Extraction**: Developed a method to preserve learned relations using spectral analysis and class-wise regularization, achieving SOTA on FewRel and TACRED datasets
  - **Transitioning Representations between Languages for Cross-lingual Event Detection via Langevin Dynamics**: Created a method using Langevin Dynamics and semantic preservation to improve target-language performance, achieving state-of-the-art results

Projects

Vistral-7B-Chat *Vistral*

- Introduced Vistral-7B-chat, a SOTA conversational Large Language Model for Vietnamese. Outperforms ChatGPT and Gemini in Vietnamese benchmarks. Achieved over 100,000 downloads
- Tools Used: HuggingFace, PyTorch

CUDA 101 Series: GEMM Optimization *gemm-101*, *hopper-gemm-101*

- Implemented GEMMs from scratch, progressively optimizing to **outperform** the cuBLAS performance for both float and half on both **Ampere** and **Hopper** GPUs
- gemm-101: Focused on basic CUDA core optimizations, improving memory access patterns through coalescing, bank conflict, vectorized memory access, and multi-level tiling (thread, warp, block)
- cute-gemm-101: Explored Tensor Core optimizations, basic Cutlass CuTe concepts including CuTe Algebra, CuTe Swizzle, Asynchronous Operation, Barrier, and Pipelining
- hopper-gemm-101: Implemented GEMMs using Cutlass CuTe with Hopper new features: Tensor Memory Accelerator (TMA), Warp Group Matrix Multiply-Accumulate (WGMMA). Improved performance with advance techniques: Warp Specialization, Persistent Kernel.
- Tools Used: CUDA, C++, Cutlass CuTe, NVIDIA Nsight Compute, Python

CUDA 101 Series: Flash Attention *flash-attn-101*, *cute-flash-attn-101*

- Developed simplified implementations of Flash Attention from scratch for self-learning purposes: Implemented naive attention, CUDA core Flash Attention 1 (FA1) and 2 (FA2), and Tensor core FA2 using Cutlass CuTe, inspired by Tridao's implementation, and added detailed comments for learning purposes
- Provided Python bindings for the Tensor core FA2 implementation
- Tools Used: CUDA, C++, Cutlass CuTe, NVIDIA Nsight Compute, Python

CUDA 101 Series: Hopper (Work in Progress)

- Ongoing project exploring NVIDIA Hopper architecture features for self-learning purposes
- Planning to apply these concepts to implement efficient fused transformer layers
- Tools Used: CUDA, C++, Cutlass CuTe, NVIDIA Nsight Compute, Python

Open Source Contributions

- **Unsloth**: Implemented a faster rotary embedding kernel using Triton
  - Pull Request: #238
  - Tools Used: Triton

Publications

Taipan: Efficient and Expressive State Space Language Models with Selective Attention Arxiv
Chien Van Nguyen, ***Huy Huu Nguyen***, Trung Bui, Viet Dac Lai, Franck Dernoncourt, Thien Huu Nguyen

A Spectral Viewpoint on Continual Relation Extraction EMNLP 2023 Findings
Huy Nguyen, Chien Nguyen, Linh Ngo, Anh Luu, Thien Nguyen

Transitioning Representations between Languages for Cross-lingual Event Detection via Langevin Dynamics: EMNLP 2023 Findings
Chien Nguyen, ***Huy Nguyen***, Franck Dernoncourt, Thien Nguyen

Technologies

Languages: C/C++, Python, CUDA, Triton, PTX

Technologies: PyTorch, HuggingFace, Cutlass CuTe, NVIDIA Nsight Compute, CMake

Academic Service

Conference Reviewer: ACL ARR 2023, ACL ARR 2024

Awards

Prestigious Recognition Academic Excellence

- Lokey Award (2024-25)
  - Prestigious award from the Department of Computer Science, University of Oregon
  - Granted to top graduate students in science fields
  - Funded by Mr. Lorry I. Lokey, allocated by UO Division of Graduate Studies

First Place Award AI Competitions

- BKAI-NAVER Challenge 2022 (May 2022)
  - Task: Intent Detection and Slot Filling for smarthome systems