Wentao Ni

+1 (858) 250 9583 w2ni@ucsd.edu wennitao

Research Interest

My research interest mainly lies in machine learning system, high performance computing, programming languages and computer architecture. Currently I am interested in architectures and systems accelerating machine learning and LLM. I want to contribute to faster computations and less resources needed in machine learning and other computation tasks.
I am also interested in programming languages, especially compilers. I am fascinated by the aggressive optimization tricks and how to ensure the codes run right.

Education

- **University of California San Diego** San Diego, CA
  PhD, Computer Science and Engineering *Sep. 2024 – Present*
- **Shanghai Jiao Tong University** Shanghai, China
  B.S. in Computer Science and Technology, ACM Class, Zhiyuan College; GPA: 86.1/100.0 *Sep. 2020 – June. 2024*

Publications

- **JUNO: Optimizing High-Dimensional Approximate Nearest Neighbour Search with Sparsity-Aware Algorithm and Ray-Tracing Core Mapping**
  Zihan Liu, **Wentao Ni**, Jingwen Leng, Yu Feng, Cong Guo, Quan Chen, Chao Li, Minyi Guo, Yuhao Zhu
  ASPLOS 2024.

Research Experience

- **Optimize High-Dimensional ANNS with Ray-Tracing Core** Shanghai Jiao Tong University
  Undergraduate Researcher, advised by Prof. Jingwen Leng *June 2022 - May 2023*
  - We study the inefficiency of the typical IVFPQ pipeline and identify sparsity and spatial similarity in codebook usage.
  - We design a threshold-based selective algorithm to rapidly filter out the unnecessary search points leveraging the sparsity and spatial locality and propose a mapping for our algorithm to run on the RT core.
  - We study how to generalize the existing kNN-RT core mapping to ANN search with arbitrary dimensions, in aspects of approximation method, metrics and system design, and propose JUNO, an end-to-end high-dimensional ANN search engine with both algorithmic enhancement and optimized hardware mapping.

- **Combining Graph & Tensor Transformation with Scheduling via Compilers** Duke University
  Research Assistant, advised by Prof. Yiran Chen *Sep. 2022 - Dec. 2023*
  - State Space Models (SSM) are known for its capability in long range modeling. However, the authors needed to carefully design CUDA kernels for SSMs.
  - We explore TVM and EinNet to combine graph-level and tensor-level transformation on tensor expressions, along with the scheduling space. We implement it based on TVM te expression.
  - There are opportunities for aggressive kernel fusion of memory-bound operators, which remains a future research topic.

- **Search Efficient Network Architectures for LLM with Linear Complexity** Duke University
  Research Assistant, advised by Prof. Yiran Chen *Sep. 2022 - Dec. 2023*
  - We consider using NAS-based distillation to search architecture computing attention in linear complexity, like RetNet and Hyena Hierarchy.

- Previously, we carefully study State Space Models (SSM) and its application in neural networks. We find it efficient in both memory and computation, so we want to find ways to apply it to transformers.

- **Improve Quantization by Searching Configurations** Microsoft Research Asia
  Intern, advised by Zhenhua Han, Yuqing Yang *Sep. 2023 - Dec. 2023*
  - We have previous works studying efficient kernels for sparse matrix computation (PIT and sparTA). So we consider building a search space of quantization configurations, such as quantization method, granularity, number of bits etc. With the help of sparse kernels, hybrid quantization can be perfectly supported.

Projects

- **Mx* Compiler**
  - A Compiler from Mx* language (C++ like) to RV32I Assembly, with static optimizations on LLVM IR, implemented in Java.
  - Front-end parser and lexer use Antlr library. Middle-end LLVM IR generation, back-end RV32I assembly codegen, static optimizations are all handwritten.
  - Static optimizations include Static Single-Assignment Form (SSA), Sparse Conditional Const Propagation (SCCP), Aggressive Dead Code Elimination (ADCE) and loop inline.
- **RISC-V CPU**
  - A Tomasulo RISC-V CPU with iCache and branch predictor with 2-bit saturating counter, implemented in Verilog RTL.
- **Pytorch-like library**
  - Follow the course 10-414/714: Deep Learning System Course, CMU.
  - Implemented two levels. Python-level: forward and backward computation of operators, nn.modules, weight initializations, optimizers, dataloaders etc. Numpy-level: NDArray (handling memory, shape, strides and offset) with both CPU and CUDA backend.

Awards and Honors

- **Mathematical Contest In Modeling** 2021
  ***Outstanding Winner**, one of 36 outstanding winners out of 26112 teams*
- **The 45^{th} ICPC Yinchuan Regional Programming Contest** 2021
  ***Gold Medal**, Rank: 19/491* *Yinchuan, China*
- **CCPC 2020 Weihai Regional Programming Contest** 2020
  ***Gold Medal**, Rank: 18/384* *Weihai, China*
- **The 45^{th} ICPC Nanjing Regional Programming Contest** 2020
  ***Gold Medal**, Rank: 9/548* *Nanjing, China*
- **SJTU Zhiyuan Honors Program Scholarship** 2020, 2021
- **SJTU ACM Class Scholarship** 2020
- **National Olympiad in Informatics** 2019
  ***Bronze Medal**, Rank: 188/263* *Guangzhou, China*
- **National Olympiad in Informatics Winter Camp** 2019
  ***Silver Medal**, Rank: 152/381* *Guangzhou, China*

Part-time Working Experience

- **Teaching Assistant** 2022 - 2023
  Shanghai Jiao Tong University *Shanghai, China*
  - Data Structure (CS1951) Mar. – June 2022
  - Principle and Practice of Computer Algorithms (CS1952) June – Aug. 2022
  - Advanced Compiler (CS2965) Mar. – June 2023
    * Design a coding assignment about loop transformation on Polyhedral, based on SCoP format.