Luning Wang

EDUCATION

Tsinghua University **Beijing, China**

Department of Electronic Engineering

B.E. in Electronic Information Science and Technology ***08/2020- 06/2024***

- Overall GPA: 3.76/4.0
- GRE: Verbal 155/ Quantity 170, Total 325, AW 4.0
- TOEFL: Reading 29/ Listening 26/ Speaking 22/ Writing 25/ Total 102
- Relevant Courses: Signals and Systems(4.0), Media and Cognition(4.0), Communications and Networks(4.0), Digital Image Processing(4.0), Fundamental of digital logic and processor(3.6), Database(3.6)

The University of Hong Kong **Hong Kong**

Musketeers Foundation Institute of Data Science

Undergraduate Research Assisstant, Advisor: Dr. Xihui Liu ***07/2023- 09/2023***

RESEARCH INTERESTS

- Deep Learning (CV & NLP)
- Signal Processing
- Biomedical Engineering

PUBLICATIONS

Shiyao Li, Xuefei Ning, Ke Hong, Tengxuan Liu, **Luning Wang,** Xiuhong Li, Kai Zhong, Guohao Dai, Huazhong Yang, Yu Wang. **"LLM-MQ: Mixed-precision Quantization for Efficient LLM Deployment"**. *The Efficient Natural Language and Speech Processing Workshop with NeurIPS 2023*

RESEARCH EXPERIENCES

NICS Lab, Energy Efficient Computing Group (Tsinghua University) **Beijing, China**

Nanoscale Integrated Circuits and System Lab, Energy Efficient Computing Group (NICS-EFC) is leaded by Professor Yu Wang, in Electronic Engineering Department, Tsinghua University. The group is committed to the research of energy-efficient circuits and systems design methodology towards the Artificial Intelligence (AI) scenario: Multi-agent Reinforcement Learning Algorithm, Efficient and Robust DL system, Domain Specific Acceleration, and Multi-agent system.

Undergraduate Research Assistant, Advisor: Prof. Yu Wang ***03/2023- 09/2023***

- **Project: Low-Bit Quantization with Mixed Precision for Trillion-Parameter-Level Large Language Models**
  - Conducted sensitivity tests on leading Large Language Models (e.g., OPT, LLaMa), gathering per-block and per-layer sensitivity data and generating informative charts to guide subsequent mixed-bit quantization strategies.
  - Assisted the Ph.D student in experimental evaluation of our grouping and reordering quantization strategy. Evaluated quantization losses under various reorder strategies and assessed losses for automatic and heuristic bit-width allocation, finally achieving an average bit-width of 2.8 bits.
  - Contributed to the writing of section "Introduction" and "Method" in our research paper, which was accepted by the Efficient Natural Language and Speech Processing Workshop with NeurIPS 2023.

Undergraduate Research Assistant, Advisor: Prof. Yu Wang ***10/2022- 02/2023***

- **Project: Detection Task for Small Targets in Simulated Combat Systems**
  - Scraped raw small target datasets (e.g., ships, tennis balls) from the web and conducted data preprocessing tasks, including data cleaning and file structure conversion, to create datasets adhering to the standard format of the YOLO open-source framework.
  - Utilized object detection frameworks such as SSD, YOLOv3, YOLOv5, and YOLOv7 for training and validation on the processed small target dataset. Achieved mAP-50 value exceeding 0.90 on the validation set for YOLOv5 and YOLOv7

models.

- Produced a video demo showcasing object detection results using the OpenCV library and further developed user-friendly interface code, enabling end-to-end processing of input videos and outputting detection box results.

Musketeers Foundation Institute of Data Science (HKU) **Hong Kong**

HKU Musketeers Foundation Institute of Data Science (HKU IDS) is the University's pioneering research institute, connecting experts from various fields, including those recruited in the 2021 HKU-100 Recruitment Campaign. It aims to become a global leader in data science research, fostering collaborations with partners worldwide.

Undergraduate Research Assistant, Advisor: Dr. Xihui Liu ***07/2023- 09/2023***

- **Project: Research on the Evaluation of Quantization Loss in Quantized Models**
  - Assisted in researching on current metrics, datasets and tasks for Large Language Models' evaluation. Be responsible for developing an enhanced, versatile evaluation framework for Large Language Models by extending the lm-evaluation-harness open-source project.
  - Wrote code to integrate some latest datasets (e.g. MMLU, CEVAL) and performed necessary encapsulation.
  - Evaluated quantization loss on the SqueezeLLM quantization framework across multiple tasks. Conducted testing on over 20 tasks hosted on Hugging Face, gathered data using various metrics, and generated charts to analyze quantization loss differences across different task categories.
  - Conducted a fine-grained analysis of the task 'sst' to examine the specific data features responsible for the highest quantization loss(more than 50%). Compiled statistics on the ratio of false positives to false negatives, overlap between misclassified samples in the original and quantized models, and other factors to analyze the specific causes of high quantization loss, ultimately revealing it was the destruction of FFN layers that led to that loss.

INTERNSHIP EXPERIENCES

ByteDance **Beijing, China**

Established in 2012, ByteDance entered the scene with Toutiao in August 2012 and followed up with Douyin in September 2016. In 2017, it made a significant global impact by launching TikTok and later merging with Musical.ly. Today, ByteDance's TikTok is a dominant global platform for short-form mobile videos, contributing to its remarkable success in the mobile internet market.

Algorithm Intern, Advisor: Ning Wang ***09/2023- 01/2024***

- **Project: The Development of an Appeal Chatbot for TikTok Moderation System**
  - Developed the retrieval-augmented generation (RAG) component for the chatbot, employing technologies such as SBert Encoder and FAISS. This encoder-retriever segment attained a top-1 recall rate exceeding 50% on the OpenBookQA dataset.
  - Employed our RAG pipeline to enhance the generation of Large Language Models including baichuan, mistral, and gpt-3.5-turbo for QA tasks, and achieved an improvement of 10% ~ 20% in accuracy on the OpenBookQA dataset.
  - Contributed to the development of the explanation generation model for reasoning. Implemented tuning strategies, including supervised fine-tuning and in-context learning with curated prompts, attaining an F1 score surpassing 70% in identifying violations within TikTok's moderation data.

SKILLS

- **Programming Languages**：Proficient in Python, C/C++, Matlab. Have fundamental knowledge of C#, Verilog, HTML, CSS, JavaScript, SQL, etc.
- **Software Tools**：Proficient in Linux, Git, Latex, MySQL, etc.

HONORS&AWARDS

- Comprehensive Excellence Scholarship of Tsinghua University (Top 30% in major) ***2022-2023***
- Tsinghua Friends-Toyota Scholarship (5000 CNY) ***2022-2023***
- First Prize in the 5th 'Huiye Cup' Software Design Competition (5000 CNY) ***2021-2022***
- Volunteer Excellence Scholarship of Tsinghua University (2000 CNY) ***2021-2022***