Zheng WANG

Email: zwang3478@gatech.edu | Tel: (+1) 404 717 4045 | Birth Date: 03/22/2003 | Website: https://zkbig.github.io/

EDUCATION BACKGROUND

Georgia Institute of Technology (Gatech), USA. 08/2023 – present
M.Sc. in Computational Science and Engineering (CSE) CoC Home Unit | Current GPA: 4.0/4.0

Beijing University of Technology (BJUT), Beijing & University College Dublin (UCD), Ireland. 09/2019 – 07/2023
B.Eng. in Internet of Things | GPA: 3.85/4.2 | Ranking: 2/57 | First-class Degree & Honors Degree

SELECTED PUBLICATIONS

- **ZoomVLM: A Tuning-Free Framework for Efficient Video Understanding via Adaptive Zooming in Vision-Language Models**
  Zhongzhi Yu*, **Zheng Wang***, Zhenyang Chen, Chaojian Li, Hyewon Suh, Yonggan Fu, Dachuan Shi, Hongxu Yin, Jan Kautz, Pavlo Moclchanov, Yingyan (Celine) Lin
  Under review, ICLR 2024.

- **Model Tells You Where to Merge: Adaptive KV Cache Merging for LLMs on Long-Context Tasks**
  Zheng Wang, Boxiao Jin, Zhongzhi Yu, Minjia Zhang
  Under review, ICLR 2024.

- **Unveiling and Harnissing Hidden Attention Sinks: Enhancing Large Language Models without Training through Attention Calibration**
  Zhongzhi Yu*, **Zheng Wang***, Yonggan Fu, Huihong Shi, Khalid Shaikh, Yingyan (Celine) Lin
  International Conference on Machine learning (**ICML**), 2024

- **When Lienar Attention Meets Autoregressive Decoding: Towards More Efficient and Linearized Large language Models**
  Haoran You, Yichao Fu, **Zheng Wang**, Amir Yazdanbakhsh, Yingyan (Celine) Lin
  International Conference on Machine learning (**ICML**), 2024

- **EDGE-LLM: Enabling Efficient Large Language Model Adaptation on Edge Devices via Unified Compression and Adaptive Layer Voting**
  Zhongzhi Yu*, **Zheng Wang***, Yuhan Li, Haoran You, Ruijie Gao, Xiaoya Zhou, Sreenidhi Reedy Bommu, Yang Katie Zhao, Yingyan Celine Lin
  Design Automation Conference (**DAC**), 2024

- **XRouting: Explainable Vehicle Rerouting for Urban Road Congestion Avoidance using Deep Reinforcement Learning**
  Zheng Wang, Shen Wang
  IEEE International Smart Cities Conference (**ISC2**), 2022

SELECTED RESEARCH EXPERIENCES

Explainable Vehicle Rerouting for Urban Road Congestion Avoidance via Deep Reinforcement Learning (UCD)
***Research Assistant* | Advisor: Prof. Shen Wang** 11/2021 – 06/2022
Designed and implemented a dynamic vehicle rerouting system for urban congestion avoidance named **XRouting** by integrating a policy-based DRL algorithm (PPO) and the revised gated transformer (GTr) architecture, demonstrating superior training stability, computational efficiency and convergence rate.

Efficient LLM Adaptation via Layerwise Unified Compression and Adaptive Tuning & Voting (Georgia Tech)
***Research Assistant* | Advisor: Prof. Yingyan (Celine) Lin** 09/2023 –12/2023
Implemented a unified LLM compression method to reduce computation cost, offering cost-effective layer-wise pruning ratios and quantization bit-precision policies. Implemented a memory-efficient training pipeline for LLM that select a subset of layers during each iteration and then adaptively combines their outputs for the final evaluation, thus reducing backpropagation depth and memory overhead while maintain competitive performance.

Unveiling and Harnissing Hidden Attention Sinks: Enhancing LLMs through Attention Calibration (Georgia Tech)
***Research Assistant* | Advisor: Prof. Yingyan (Celine) Lin** 12/2023 –03/2024
Comprehensively explored and analyzed attention sinks in LLMs. Designed and implemented a training-free **Attention**

Calibration (ACT) technique that automatically optimizes the attention distributions on the fly during inference in an input-adpative manner. ACT can enhance the accuracy of a wide array of pretrained LLMs up to 3.16% across various tasks.

Model Tells You Where to Merge: Adaptive KV Cache Merging for LLMs on Long-Context Tasks (UIUC)
***Research Intern* | Advisor: Prof. Minjia Zhang** 05/2024 –08/2024
Designed and implemented a dynamic KV cache merging approach **KVMerger,** which can adaptively merge key and value states for LLMs via Gaussian Kernel weighted method to compress KV cache while maintaining competitive performance on long-context tasks. KVMerger also outperms the KV cache eviction methods for Group-Query-Attention based LLMs.

SELECTED AWARDS AND HONORS

Excellent Graduates of Bejing	06/2023
Presidential Fellowship in 2021-2022 Academic Year (8 places for all students)	11/2022
Xiaomi Special Scholarship in 2021-2022 Academic Year (10 places for all students)	11/2022
1st Prize of the China Undergraduate Mathematical Contest in Modeling, Beijing District	11/2020
2nd Prize of the China Undergraduate Mathematical Contest in Modeling, Undergraduate Group (2.3%)	11/2020
2nd Prize Award of the 1st International Competition on Intelligent Simulation of Transport Infrastructure	03/2022
2nd Prize of the 7th China College Students' Internet+ Innovation and Entrepreneurship Competition	08/2021
Innovation and Entrepreneurship Award of BJUT in 2020-2021 Academic Year	12/2021
Outstanding Student Leaders of BJUT	2020-2022
Merit Student Award of BJUT	2020-2022
Learning Excellence Award of BJUT	2020-2022

TEACHING

CSE 8803 Machine Learning for Neural and Behavior Data 2024 Fall
Teaching Assistant, Georgia Tech | Instructor: Anqi Wu

SKILLS

Programming Languages: Python | C | Java | MATLAB| SQL| Verilog HDL
Libraries: PyTorch | TensorFlow | Ray | RLlib | Gym | Flow | Scikit-Learn | NumPy | SciPy | Pandas
Software: SUMO | LTspice | Visual Studio | PyCharm | Eclipse | IntelliJ | LaTeX | Quartus | ModelSim | Keil uVision5 | EdSim51 | Linux | Wireshark | Packet Tracer | Microsoft Office