Yushi Huang

20376156@buaa.edu.com · Harahan

Education

Beihang University (BUAA) | Shen Yuan Honors College 2020.09 – now
(Bachelor of *Computer Science and Technology*) **GPA:** 3.87/4.00 (93.2/100) **Rank:** 15/265

Experience

Model Deployment & Inference (AI Service of Deploy, SenseTime) 2023.05 – 2023.08
- Optimize some specific operators for CV models.
- Optimize inference time and memory cost for the Stable Diffusion model.

Quantization on Diffusion Models (Efficient Model RD, SenseTime) 2023.09 – 2023.11
Supervisor: Ruihao Gong
- Implement some SOTA algorithms, like Q-Diffusion, PTQD...
- Explore novel low-bit quantization methods on diffusion models according to their temporal properties and implement an end-to-end quantization pipeline.

Large Language Model Quantization Benchmark (Efficient Model RD, SenseTime) 2023.12 – 2024.03
Supervisor: Ruihao Gong
- Implement many quantization methods for LLM, like SpQR, GPTQ, SmoothQuant, OmniQuant...
- Participate in building an end-to-end LLM quantization tool, which integrates a variety of quantization algorithms and supports multiple inference backends, including tensorRT-LLM, LightLLM...
- Explore key challenges and best practices for quantization on LLM under different conditions.

Publications

Authors who equally contributed to a publication are marked with a *.

1. **TFMQ-DM: Temporal Feature Maintenance Quantization for Diffusion Models**
   Yushi Huang*, Ruihao Gong*, Jing Liu, Tianlong Chen, Xianglong Liu
   IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024. **(Highlight, Acceptance rate: 2.8%)**
2. **LLM-QBench: A Benchmark Towards the Best Practice for Post-training Quantization of Large Language Models**
   Ruihao Gong*, Yang Yong*, Shiqiao Gu*, **Yushi Huang***, Yunchen Zhang, Xianglong Liu, Dacheng Tao
   In submission to Conference on Language Modeling (COLM), 2024.
3. **PTSBench: A Comprehensive Post-Training Sparsity Benchmark Towards Algorithms and Models**
   Zining Wang, Jinyang Guo, Yang Yong, Ruihao Gong, Aishan Liu, **Yushi Huang**, Jiaheng Liu, Xianglong Liu
   In submission to ACM International Conference on Multimedia (ACM MM), 2024.

Honors & Awards

Beihang University 2020 Academic Excellence Award–First Prize (Top 8%) 2021.10
2021 National College Students Mathematics Competition–Second Prize (Top 10%) 2021.12
2021 National College Students Physics Competition (Some Regions)–Second Prize (Top 10%) 2022.12
Beihang University 2021 Academic Excellence Award–First Prize (Top 8%) 2022.10
Beihang University 2022 Academic Excellence Award–Second Prize (Top 15%) 2023.10

Skills

- **Programming Languages**: Python, C, Java
- **Scientific Packegs**: Pytorch, Numpy

Others

- **Research Interests**:
  - Model compression, including model quantization, model pruning...
  - Model efficiency for LLM and Diffusion models.
- **Languages**:
  - Mandarin Chinese (native)
  - English: 107 (R: 28 L: 29 S: 23 W: 27) in TOEFL iBT TEST