Geon Park

daniel-geon-park | username | geon.park@kaist.ac.kr | +82.10.4290.0250

Summary

I am a Ph.D. student at Machine Learning and Artificial Intelligence (MLAI) lab in KAIST, South Korea, under the supervision of Prof. Sung Ju Hwang.

My research interests include:

- Efficient Deep Learning Model Inference
  - Sparse Attention, Long Context LLM
  - Weight Quantization, Pruning
  - Neural Architecture Optimization
  - Writing Efficient CUDA Kernels for Deep Learning
- LLM Agents
- Natural Language Processing

Education

2023 - present	PhD student at **KAIST** Graduate School of Artificial Intelligence	
2021 - 2023	**KAIST** Master's Degree in Artificial Intelligence	(GPA: 2.43/4.3)
2017 - 2021	**Sogang University** Bachelor's Degree in Computer Science	(GPA: 3.86/4.3)
2018	**Pennsylvania State University** Student Exchange Program	

Pre-prints

Lee, Heejun et al. (Feb. 2025). *InfiniteHiP: Extending Language Model Context Up to 3 Million Tokens on a Single GPU*. arXiv:2502.08910 [cs]. URL: http://arxiv.org/abs/2502.08910.

Publications

Jeong, Wonyong et al. (2021). "Task-adaptive neural network search with meta-contrastive learning". In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 21310–21324.

Yoon, Jaehong et al. (2022). "Bitwidth heterogeneous federated learning with progressive weight dequantization". In: *International Conference on Machine Learning*. PMLR, pp. 25552–25565.

Park, Geon et al. (2023). "BiTAT: Neural Network Binarization with Task-Dependent Aggregated Transformation". In: *Computer Vision – ECCV 2022 Workshops*. Vol. 13807. Cham: Springer Nature Switzerland, pp. 50–66.

Lee, Heejun et al. (2024). "A Training-Free Sub-quadratic Cost Transformer Model Serving Framework with Hierarchically Pruned Attention". In: *The Thirteenth International Conference on Learning Representations*.

Kim, Kangsan et al. (2025). "VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3295–3305.

Work Experience

Internship at AITRICS Sep 2020 - Mar 2021

I developed TANS (Task-Adaptive Neural Network Search) during internship at AITRICS.

Internship at DeepAuto Jun 2023 - present

I developed HiP Attention (A Training-Free Sub-quadratic Cost Transformer Model Serving Framework with Hierarchically Pruned Attention) and integrated it with SGLang framework during internship at DeepAuto.