Last updated in September 2025

Heejun Lee

Seoul, South Korea | ainl@kaist.ac.kr | +82 10-7757-5176 | https://github.com/gmlwns2000

Professional Summary

Ph.D. student at KAIST specializing in efficient deep learning and large language models. Proven track record of developing novel sparse attention mechanisms with multiple first-author publications at top-tier conferences (ICLR). Passionate about translating cutting-edge research into cost-effective, high-performance AI products.

Education

Korea Advanced Institute of Science and Technology, Combined M.S./Ph.D. in Artificial Intelligence — Sep 2024 – Feb 2029 (Expected)

Korea Advanced Institute of Science and Technology, BS in Computer Science — Mar 2020 – Aug 2024

- GPA: 3.97/4.3
- College of Engineering Dean's List (Spring 2022)
- College of Engineering Leadership Award on Research Excellence (Spring 2022, Spring 2023)

Experience

AI Research Engineer, DeepAuto.ai – Seoul, South Korea — Dec 2023 – Present

- Developed **ScaleServe**, a cost-efficient LLM serving framework that **reduces end-to-end serving costs by approximately 52%** by integrating novel, training-free attention mechanisms.
- Invented **HiP Attention (ICLR 2025)**, a training-free attention algorithm that **speeds up long-context inference by 50%** and enables serving million-token contexts on a single GPU via KV cache offloading.
- Designed **Delta Attention**, a novel correction algorithm that **boosts sparse attention accuracy by 20-30%** on the RULER benchmark with only a marginal (<10%) latency overhead.
- Engineered and integrated custom attention modules into serving frameworks like vLLM and SGLang, reducing computational complexity for long contexts from quadratic ($O(n^2)$) to near-linear ($O(n)$).

Publications

** Denotes equal contribution*

Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction — arXiv Preprint

Jeffery Willette, ***Heejun Lee***, Sung Ju Hwang
(Github)

InfiniteHiP: Extending Language Model Context Up to 3 Million Tokens on a Single GPU — arXiv Preprint

Heejun Lee*, Geon Park*, Jaduk Suh*, Sung Ju Hwang
(Github)

A Training-Free Sub-quadratic Cost Transformer Model Serving Framework... — ICLR 2025

Heejun Lee*, Geon Park*, Youngwan Lee*, Jaduk Suh*, et al.
(Github)

Training-Free Exponential Extension of Sliding Window Context with Cascading KV Cache — ICLR 2025

Jeffrey Willette, ***Heejun Lee***, Youngwan Lee, Myeongjae Jeon, Sung Ju Hwang
(Github)

SEA: Sparse Linear Attention with Estimated Attention Mask — ICLR 2024

Heejun Lee, Jina Kim, Jeffery Willette, Sung Ju Hwang
(Github)

Sparse Token Transformer with Attention Back Tracking ICLR 2023

Heejun Lee, Minki Kang, Youngwan Lee, Sung Ju Hwang
(Github)

Skills

Languages: Python, C++, C#
Frameworks & Libraries: PyTorch, Hugging Face, vLLM, SGLang, OpenAI Triton, .NET