<Professor Insu Han's Lab.>

Flexible Machine Learning Algorithm Lab

■ **Contact information**

Professor : Insu Han TEL : 7477

Lab. : Flexible ML Algorithm Lab

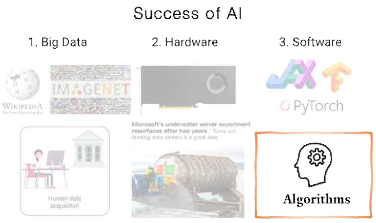
Website : flexml.kaist.ac.kr / insuhan.github.io

■ **Current state of the Lab. (in 2025 Spring Semester)**

2 incoming Master's Student, 8 undergraduate students

■ **Research Areas**

Our research group focuses on developing scalable algorithms for large-scale machine learning problems, combining practical efficiency with rigorous theoretical analysis. We tackle challenges at the forefront of AI and machine learning, aiming to improve the performance and applicability of cutting-edge models and techniques.



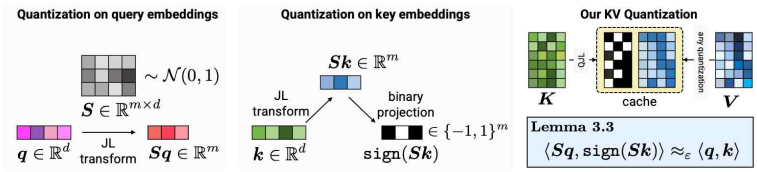
Several research directions are

1. Development of large-scale machine learning approximation algorithms

2. Development of quantization algorithms for inference foundation models

3. Development efficient algorithms for genomic foundation models for cancer classification

3. Acceleration of core operations in foundation models through matrix sparsity approximation



■ **Recommended courses & Career after graduation**

- **Programming structures (EE209, EE309, CS101, CS109, etc)**
- **Data structure and algorithm (EE205, CS206, CS300, etc)**
- **Machine learning (EE331, EE412, EE424, CS475, etc)**
- **Mathematics (EE210, MS250, MS350, MS355, MS365, etc)**
- **Industrial experiences before/after graduate school**

■ **Introduction to other activities besides research**

- **Physical activities that achieve progressive goals, such as hiking, cycling, swimming**

■ **Introduction to the Lab.**

Established in September 2024, the lab offers you mentorship from highly motivated young professor working on cutting-edge researches, and the unique opportunity to become one of the first alumni of the group.

■ **Recent research achievements ('22~'25)**

- **PolarQuant: Quantizing KV Caches with Polar Transformation, Under Review 2025**
- **BalanceKV: KV Cache Compression through Discrepancy Theory, Under Review 2025**
- **CalibQuant: 1-Bit KV Cache Quantization for Multimodal LLMs, ICML 2025 LCFM workshop**
- **QJL: 1-Bit Quantized JL Transform for KV Cache Quantization with Zero Overhead, AAAI, 2025**
- **Cell2Sentence: Teaching Large Language Models the Language of Biology, ICML 2024**
- **HyperAttention: Long-context Attention in Near-Linear Time, ICLR 2024**
- **Near Optimal Reconstruction of Spherical Harmonic Expansions, NeurIPS 2023**
- **KDEformer: Accelerating Transformers via Kernel Density Estimation, ICML 2023**
- **Fast Neural Kernel Embeddings for General Activations, NeurIPS 2022**
- **Scalable MCMC Sampling for Nonsymmetric Determinantal Point Processes, ICML 2022**
- **Random Gegenbauer Features for Scalable Kernel Methods, ICML 2022**