Ph.D. Jemin Lee, leejaymin@etri.re.kr, Homepage
Electronic Telecommunication Research Institute (ETRI), Republic of Korea

RESEARCH EXPERIENCE

AI Model Compression and Evaluation (ETRI and UST) Dec. 2018 – Present
Senior Researcher

- Designed and deployed an integrated evaluation framework that measures accuracy deltas caused by post-training quantization on instruction-tuned LLMs up to **405B** parameters in a Ray + vLLM multi-node cluster.
- Specialized in compressing diverse model families — CNNs, (Hybrid) Vision Transformers, and LLMs — for both cloud and edge targets.
- Post-training quantization: identified optimal calibration parameters, crafted reconstruction-error-aware quantizers to maximize post-compression accuracy, and developed mixed non-linear quantization which enabled fully integer-only ViT inference via function approximation of GELU and Softmax with no retraining.
- Quantization-aware training: developed mixed non-linear quantization for ViT models using QAT, assigning optimal methods per non-linear op (LayerNorm, Softmax, GELU) based on SQNR sensitivity analysis.
- Mixed-precision methodology: selected precision per-layer/operation to trade off compute throughput versus model quality.

AI Compiler and Kernel Optimization (ETRI and UST) Dec. 2018 – Present
Senior Researcher

- **Fully-INT8 FlashAttention Triton kernel:** implemented softmax and all intermediate ops entirely in the integer domain via scale-offset quantisation, achieving FP-free attention inference with Triton-lang.
- Continuously authoring Triton-lang kernels for efficient attention variants across multiple NVIDIA and AMD GPU architectures.
- ONNX-level and high-level IR optimisation (Glow / TVM Relay / PT 2.0 Dynamo) enabling new ops and improving mapping to HW accelerators.
- Auto-scheduler development (CPrune, ACLTuner, ML^{2}Tuner, and Luthier) and co-optimisation with Ansor to generate highly efficient kernels on DragonBoard 865, ODroid, Edge-R and so on.
- Extended Apache TVM's VTA backend and built an **EVTA CodeGen** path targeting an enhanced accelerator with widened GEMM pipelines and INT4/INT8 mixed-precision support.

Energy Aware Mobile Computing (CNU and KAIST) Period: Apr. 2011 – Oct. 2018
Postdoctoral Researcher

- Developed core engines for SuggestBot, focusing on context-based association/suggestion applications: collected in-the-wild data from conversation-based interactions, mobile/wearable sensors.
- Engineered software for collecting sensor and interaction data from mobile and wearable devices.
- Improved energy efficiency of continuous sensing via data fusion and inference.

WORK EXPERIENCES

Department of Artificial Intelligence, UST Daejeon, South Korea
Assistant Professor Sept. 2023 – Present
Efficient Computing Laboratory

AI Research Laboratory, ETRI Daejeon, South Korea
Senior Researcher Dec. 2018 – Present

Interactive Computing Laboratory, KAIST Daejeon, South Korea
Post-doctoral Researcher (including compulsory military service period) Nov. 2017 – Aug. 2018
Supervisor: Uichin Lee

EDUCATION

Chungnam National University *Sept. 2011 – Aug. 2017*
Ph.D. in Department of Computer Science and Engineering
Embedded System Laboratory
Thesis: Power Modeling, Analysis, and Optimization for Mobile Devices
Advisor: Hyungshin Kim
Outstanding Ph.D. Thesis Award (top 1 out of 115)

Chungnam National University *Mar. 2006 – Aug. 2011*
B.S. in Department of Computer Science and Engineering

PUBLICATIONS

Peer-reviewed Journals and Proceedings

† equal contribution, *corresponding author

J.15 A Survey on Inference Engines for Large Language Models: Perspectives on Optimization and Efficiency
Sihyeong Park, Sungryeol Jeon, Chaelyn Lee, Seokhun Jeon, Byung-Soo Kim, and **Jemin Lee***
In Preprint on ArXiv 2505.01658 May 3, 2025, May 2025

C.21 I-FlashAttention: Fully Integer Fused Attention for Efficient Vision Transformers **(WIP)**
Sehyeon Oh, Yongin Kwon, and **Jemin Lee***
ACM International Conference on Compilers, Architectures, and Synthesis for Embedded Systems (ESWEEK-CASES 2025), Work-In-Progress (WIP), September 30, 2025

C.20 Design Practices and Lessons from Deploying On-device Vision-Language Interaction in Robotic Guide Dogs
Jinse Kwon, **Jemin Lee***, and Yongin Kwon*
Thirteenth International Workshop on Assistive Computer Vision and Robotics (ICCV Workshop) 2025

C.19 TriPlanNet: Triangle Path Planning Network for A Variable Truss Robot with Deep Learning
Choonghan Lee, Leah Harris, Sehyeon Oh, Juhyung Cha, **Jemin Lee**, Yongin Kwon, and Andrew Jang-ho Bae
Thirteenth International Workshop on Assistive Computer Vision and Robotics (ICCV Workshop) 2025

C.18 Luthier: Bridging Auto-Tuning and Vendor Libraries for Efficient Deep Learning Inference
Yongin Kwon, Joo Hyoung Cha, Jubin Lee, Misun Yu, Jeman Park, and **Jemin Lee***
(Top Conf.) The International Conference on Compilers, Architectures, and Synthesis for Embedded Systems (CASES 2025), and ACM Transactions on Embedded Computing Systems (TECS), ACM Artifact Available & Evaluated, Sept. 29 2025 (NRF BK21+ IF 2).

C.17 Exploring the Trade-Offs: Quantization Methods, Task Difficulty, and Model Size in Large Language Models From Edge to Giant
Jemin Lee, Sihyeong Park, Jinse Kwon, Jihun Oh, and Yongin Kwon
(Top Conf.) In International Joint Conferences on Artificial Intelligence (IJCAI) Aug. 18 2025 (NRF

BK21+ IF 4, Acceptance Rate 19.3% (1,042 papers accepted out of 5,404 submitted).

C.16 Multi-Level Machine Learning-Guided Autotuning for Efficient Code Generation on a Deep Learning Accelerator
JooHyoung Cha, Munyoung Lee, Jinse Kwon, Jubin Lee, **Jemin Lee**, and Yongin Kwon
The 26th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems (LCTES) pp.134 - 145, 13 Jun 2025, (NRF BK21+ IF 2, Acceptance Rate 38% (16 papers accepted out of 42 submitted).

J.14 QuantuneV2: Compiler-based local metric-driven mixed precision quantization for practical embedded AI applications
Jeongseok Kim†, **Jemin Lee**†, Yongin Kwon, and Daeyoung Kim
In Future Generation Computer Systems (IF: 6.2, JCR23 Top 9.4%, Q1), Jan 2025

C.15 Optimizing Real-Time Object Detection in a Multi NPU Systems
Sehyeon Oh, Yongin Kwon, and **Jemin Lee***
In MDPI Sensors, Volume 25, Issue 5, pp. 1376 March 1 2025 EISSN 1424-8220 (IF: 3.4, JCR24 Top 30.92%, Q2), Mar 2025

C.14 ML^{2}Tuner: Efficient Code Tuning via Multi-Level Machine Learning Models
JooHyoung Cha, Munyoung Lee, Jinse Kwon, Jubin Lee, **Jemin Lee**, and Yongin Kwon
In Machine Learning for Systems Workshop (NeurIPS Workshop), Dec. 2024

C.13 Mixed Non-linear Quantization for Vision Transformers
Gihwan Kim†, **Jemin Lee**†, Sihyeong Park, Yongin Kwon, and Hyungshin Kim
The Fourth Workshop on Computational Aspects of Deep Learning (ECCV Workshop), Sep 2024

J.13 Q-HyViT: Post-Training Quantization for Hybrid Vision Transformer with Bridge Block Reconstruction for IoT Systems
Jemin Lee, Yongin Kwon, Jeman Park, Misun Yu, Hwanjun Song
IEEE Internet of Things Journal (IF 10.6, JCR23 Top 2.2%), Vol. 11, Issue 22, pp.36384-36396, 15 Nov. 2024

J.12 NEST-C: A Deep Learning Compiler Framework for Heterogeneous Computing Systems with AI Accelerators
Jeman Park, Misun Yu, Jinse Kwon, Junmo Park, **Jemin Lee***, and Yongin Kwon*
In ETRI Journal Vol. 46 Issue 5, pp.851-864 ISSN: 1225-6463 (IF 1.3), Oct 2024

C.12 ACLTuner: A Profiling-Driven Fast Tuning to Optimize Deep Learning Inference
Yongin Kwon, Joo Hyoung Cha, Jubin Lee, Misun Yu, Jeman Park, and **Jemin Lee***
In Machine Learning for Systems Workshop (NeurIPS Workshop), 2023

J.11 Pipelining of a Mobile SoC and an External NPU for Accelerating CNN Inference
Jinse Kwon, **Jemin Lee**, and Hyungshin Kim
IEEE Embedded Systems Letters, Vol. 16 Issue 2 pp. 150-153, June 2024 ISSN 1943-0663

J.10 PartitionTuner: An operator scheduler for deep-learning compilers supporting multiple heterogeneous processing units
Misun Yu, Yongin Kwon, **Jemin Lee**, Jeman Park, Junmo Park, Taeho Kim
ETRI Journal Vol 45 Issue 2 pp. 187-357, Apr 2023 (JCR21 IF: 1.622) ISSN: 1225-6463, doi:

https://doi.org/10.4218/etrij.2021-0446

J.9 Software-level Memory Regulation to Reduce Execution Time Variation on Multi-core Real-time Systems
Sihyeong Park, **Jemin Lee**, Hyungshin Kim
IEEE Access, Vol. 10, pp.93799-93811, Sept. 01, 2022 (JCR21 IF: 3.476) ISSN: 2169-3536

C.11 CPrune: Compiler-Informed Model Pruning for Efficient Target-Aware DNN Execution
Taeho Kim, Yongin Kwon, **Jemin Lee**, Taeho Kim, Sangtae Ha,
(Top Conf.) European Conference on Computer Vision (ECCV), pp.651–667, Oct 23-27, 2022 (NRF BK21+ IF 2, Acceptance Rate 28% (1,650 papers accepted out of 5,803 submitted).

J.8 Time-Invariant Features-Based Online Learning for Long-Term Notification Management: A Longitudinal Study
Jemin Lee, Sihyeong Park, Taeho Kim, Hyungshin Kim
Applied Sciences Vol. 12, No. 11 Article-Num. 5432, June 01, 2022 (JCR21 IF: 2.838, ISSN: 2076-3417

J.7 Quantune: Post-training quantization of convolutional neural networks using extreme gradient boosting for fast deployment
Jemin Lee*, Misun Yu, Yongin Kwon, Taeho Kim
Future Generation Computer Systems, Vol. 132, 2022, pp. 124-135 IF: 7.187

J.6 PASS: Reducing Redundant Notifications between a Smartphone and a Smartwatch for Energy Saving
Jemin Lee, Uichin Lee Hyungshin Kim
IEEE Transactions on Mobile Computing (IF 5.538, JCR20: Top 17%), Vol 19, Issue 11, 1 Dec. 2020.

J.5 Hardware Resource Analysis in Distributed Training with Edge Devices
Sihyeong Park, **Jemin Lee**, Hyungshin Kim
MDPI Electronics 2020, 9(1) 28, 26 Dec. 2019 (impact factor: 1.764)

C.10 Fire in Your Hands: Understanding Thermal Behavior of Smartphones
Soowon Kang, Hyeonwoo Choi, Sooyoung Park, Chunjong Park, **Jemin Lee**, Uichin Lee, and Sung-Ju Lee,
(Top Conf.) ACM International Conference on Mobile Computing and Networking (MobiCom) 2019 (NRF BK21+ IF 4).

J.4 Reducing Smartwatch Users' Distraction with Convolutional Neural Network
Jemin Lee, Jinse Kwon, Hyungshin Kim
Mobile Information Systems, Article ID 768954915 Mar. 2018, IF: 0.849

C.9 Analysis of Hardware Resources in Distributed Learning **(poster)**
Sihyeong Park, **Jemin Lee**, Hyungshin Kim
In Proceedings of International Workshop on Highly Efficient Neural Networks Design (co-located with EMSOFT), pp. 1-4, Seoul, South Korea, Oct. 2017.

C.8 An Ultrasound-based Indoor Localiztion Using Gaussian ASK Modulation **(WIP)**
Jinse Kwon, **Jemin Lee**, Hyungshin Kim
In Proceedings of International Conference on Indoor Positioning and Indoor Navigation, pp. 1-4, Sapporo, Japan, 18-21 Sept. 2017.

C.7 Deep Learning Training on Distributed Embedded Systems **(poster)**
Sihyeong Park, **Jemin Lee**, Hyungshin Kim

In Proceedings of the 12th lEMEK Symposium on Embedded Technology, Busan, South Korea, 18-19 May, 2017.

C.6 Extending App Pre-Launch Service with Emotion Context **(poster)**
Jinyoung Choi, **Jemin Lee**, Hyungshin Kim
In Proceedings of the 2nd ACM/IEEE International Conference on Internet-of-Things Design and Implementation (IoTDI'17) Adjunct, pp. 1-2, Pittsburgh, USA, 18-21 Apr. 2017.

J.3 QDroid: Mobile Application Quality Analyzer for App Market Curators
Jemin Lee, Hyungshin Kim
Mobile Information Systems, vol. 2016, Article ID 1740129, 11 pages, 10 Oct. 2016, IF: 1.462

C.5 Reducing Distraction of Smartwatch Users with Deep Learning
Jemin Lee, Jinse Kwon, Hyungshin Kim
In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'16) Adjunct pp. 948-953, Florence, Italy, Sept. 2016.

J.2 O-Sleep : Output-Oriented Power Saving Mode for Smartphones
Hyunwoo Joe, Jungseok Kim, **Jemin Lee**, Hyungshin Kim
Future Generation Computer Systems, 6 Jun. 2016, IF: 2.430

J.1 Automated Power Model Generation Method for Smartphones
Jemin Lee, Hyunwoo Joe, Hyungshin Kim
IEEE Transactions on Consumer Electronics, Vol. 60(2), pp. 190-197, May, 2014, IF: 1.045

C.4 Framework for automated power estimation of Android applications **(poster)**
Jemin Lee, Hyungshin Kim
International conference on Mobile systems, applications, and services (Mobisys'13), Taipei, Taiwan, pp. 541-542, Jun. 2013.

C.3 Energy Reservation Service for Smart Phone Application **(poster)**
Vincent Dupre, **Jaymin Lee**, Hyungshin Kim
3rd ACM/SIGOPS Asia-Pacific Workshop on Systems (ApSys'12) Seoul, South Korea 23-24th, July, 2012.

C.2 Smart Phone Power Model Generation Using Use Pattern Analysis
Jaymin Lee, Hyunwoo Joe, Hyungshin Kim
IEEE International Conference on Consumer Electronics(ICCE'12) Las Vegas, NV, USA 13th-16th Jan 2012.

C.1 Smartphone, where does the power go?
Jaymin Lee, Hyunwoo Joe, Hyungshin Kim
EU Korea Conference on Science and Technology (EKC'11) Paris, France, 21-23th, July 2011.

Full list of published papers: publications

STUDENT SUPERVISION

Current Students

- **Sehyeon Oh**
  MS student (Spring 2024 – Present) @ UST
  Research: AI Model Compression & Collective Communication Library (CCL)

Alumni and Former Interns

· **Kihyun Kim**
ETRI Summer Intern (2025)
Research: Quantization on RISC-V

· **Wonseok Jang**
UST Intern (2025)
Research: CI/CD and AI model inference on Jetson boards
Current Position: Graduate Student @ POSTECH

· **Sangheon Lee**
ETRI Summer Intern (2024) & UST Intern (2024)
Research: C Programming on RISC-V
Current Position: Engineer @ Hyundai Motors

· **Gihwan Kim**
ETRI Winter Intern (2024)
Research: Transformer Compression
Current Position: Graduate Student @ CNU

· **Minho Lee**
ETRI Summer Intern (2023, 2024)
Research: NPU Compiler

ACADEMIC SERVICES

Chair & Committee

· Web co-chair for ACM MobiSys 2019

· Program Committee for IeMeK 2022 2025 (대한임베디드공학회)

Board of Directors

· Editor (편집위원), Korea Information Processing Society (KIPS)
정보처리학회, 2024–Present

· Senior Board Member(상임이사), Institute of Embedded Engineering of Korea(IeMeK)
대한임베디드공학회 이사, 2022 Present

· Member of the PG601 Committee at the Telecommunications Technology Association (TTA)

External Reviewer

· Scientific Reports Aug. 2025

· Elsevier Expert Systems With Applications Aug. 2025

· WACV 2026 (2)

· Elsevier Expert Systems With Applications May 2025

· Elsevier Knowledge-Based Systems Apr. 2025

· Elsevier Expert Systems With Applications Mar. 2025

· IEEE Internet of Things Feb. 2025

· Elsevier Neurocomputing Feb. 2025

· The Journal of Supercomputing Jan. 2025

· Elsevier Neural Networks Journal 2024

· ACM CHI 2024

· CMC-Computers, Materials & Continua 2023

· Resource Efficient Deep Learning for Computer Vision 2023 (ICCV Workshop)

· MDPI Applied Science 2022 (Feb. Mar. Apr.(2))

· MDPI Electronics 2022 (Jan. Feb.)

· ACM CHI 2021

· IEEE SCC 2019
· IMWUT (UbiComp), Sept. 2018
· IMWUT (UbiComp), May 2018
· Sustainable Computing, Informatics and Systems 2018
· Journal of Medical Internet Research 2018
· IEEE Transactions on Mobile Computing 2015
· Pervasive and Mobile Computing 2014

HONORS AND AWARDS

AICompS Distinguished Paper Award 2024

· AICompS 2024.

IeMeK Best Presentation Award 2024

· Institute of Embedded Engineering of Korea 2024.

IeMeK Best Presentation Award 2022

· Institute of Embedded Engineering of Korea 2022.

KSC Best Paper Award 2019

· The Korean Institute of Information Scientists and Engineers.

IEMEK 2017 Best Presentation Award 2017

· Korean Embedded Engineering Conference 2017, Institute of Embedded Engineering of Korea.

Outstanding Ph.D. Thesis Award (top 1 out of 115) 2017

· Chungnam National University.

Embedded System Design Challenge Bronze Award (out of 28 teams) 2017

· Faster R-CNN Optimization for Embedded System, ACM SIGDA KOREA Chapter 2017.

IEMEK 2015 Best Presentation Award 2015

· Korean Embedded Engineering Conference 2015, Institute of Embedded Engineering of Korea.

KSCI 2015 Best Paper Award 2015

· Korea Society of Computer Information 2015, The Korea Society of Computer Information.

KCC 2015 Best Paper Award 2015

· Korea Computer Congress 2015, The Korean Institute of Information Scientists and Engineers.

Best Paper Award 2014

· Korea Computer Congress 2014, The Korean Institute of Information Scientists and Engineers.

Best Presentation Award 2012

· Korea Computer Congress 2012, The Korean Institute of Information Scientists and Engineers.

ISSUED PATENTS

Deep Learning Compiler with Support for Heterogeneous Computing Platforms and Its Method
Granted 03/04/2025, Korea Patent number 10-2777879
Misun Yu, Yongin Kwon, Taeho Kim, Jeman Park, **Jemin Lee**

Method and system for expecting users' mood based on status information and biometric information acquired by using user equipment

Granted 06/15/2017, Korea Patent number 10-1749706
Hyungshin Kim, **Jemin Lee**, Jinyoung Choi

Method for Detecting Indoor Zone with Bluetooth and Ultrasound of Smartphone
Granted 05/29/2017, Korea Patent number 10-1742960
Hyungshin Kim, **Jemin Lee**, Jinse Kwon

System and Method for Detecting Beacon
Granted 05/24/2017, Korea Patent number 10-1741406
Hyungshin Kim, **Jemin Lee**, Seula Hwang

Portable terminal and method for controlling a battery charging of the same
Granted 08/16/2016, Korea Patent number 10-1650038000
Hyungshin Kim, **Jemin Lee**, Donggeon Han

Search system and method of executable GUI
Granted 04/20/2015, Korea Patent number 10-1513662000
Hyungshin Kim, **Jemin Lee**, Donggeon Han

Collaborative Power Model Creation Method and Service Module With the Same
Granted 02/26/2013, Korea Patent number 10-12669710000
Hyungshin Kim, **Jemin Lee**