Chaeun Lee

1507, 113 Burim-ro, Dongan-gu, Anyang-si, Gyeonggi-do, Republic of Korea

chaeunl7765@gmail.com +82-10-9109-7765 Github GoogleScholar

Education

Korea University Institute for Continuing Education **Seoul, Republic of Korea**
Bachelor of Mathematics (45/48 credits earned) *Sep. 2021 -*

Seoul National University **Seoul, Republic of Korea**
Master of Electrical Engineering and Computer Science (Prof. Kiyoung Choi) *Mar. 2018 - Feb. 2020*
Thesis: Design and Optimization for Energy Efficient and Variation Resilient RRAM-based Neural Networks

Military Service **Republic of Korea**
Korean Army *Jan. 2012 - Jan. 2014*

Seoul National University **Seoul, Republic of Korea**
Bachelor of Electrical Engineering and Computer Science *Mar. 2011 - Aug. 2017*
Thesis: Analyzing Neural Networks System of Caenorhabditis Elegans using Modularity Algorithms

Research Interests with History

AI Simulation *Jun. 2019 -*

Experience from my first research topics of spiking neural networks ([C2]) and energy-efficient hardware implementations ([C1], [C4], [C6]) led me to another research topic which aimed at designing more accurate and faster simulation environments for analog memory-based circuit systems ([C5]).

There were three reasons that I focused on the topic: At first, it is time-consuming to build simulation environments for complicated analog-based hardware and software system, where they both have to consider continuous time domain (transient analysis). Another one is that there are no unified simulation tools with consistency. Lastly, these types of simulations require iterative methods to solve equations, which means that it takes much time to have experimental results.

To resolve the above mentioned issues, especially the third one, I focused on AI-based simulators, where AI model replace iterative methods or reduce iteration steps. Solving analog circuit equations, f, with a ton of circuit parameters for transient analysis (formulated as solving df/dt to find $f(t_i)$) requires iterative methods such as Euler's method. Based on the domain knowledge for circuit equations which is monotonic function in general, I proposed a neural architecture that find f directly in one-step ([A1], [P2]). At the time of research, as the topic was not prevalent, I delivered a tutorial at the conference ([T1]), although my work has not been published officially and been pending because of license and resource issues.

In the similar context, I have kept my attention on relevant area-physics informed machine learning, AI simulation, and generative models-in that they also find the ground truth $f(x)$ or $p(x)$ with restricted data. Among them, recently, I have opened and led a new research ([P3], [C7], [C8]) topic that a diffusion model, θ, is optimized to generate synthetic data conditioned on a discriminative vision perception model, ϕ; i.e., find $p_\theta(x|\phi)$ (or in a word, model-agnostic data generation). My interests go beyond domain-specific or task-specific simulation, and extends to AI simulation for real-world including model design and synthetic data generation.

Large-scale model *Nov. 2021 -*

During the internship at NAVER Clova, I was engaged in big model TF team in addition to OCR team. My task was finetuning a multi-task foundation model, especially for document classification as a down-stream task. More specifically, the model is composed of vision and language encoder with language decoder model. I had to finetune the model without the loss of its ability for other tasks. The results were always better than task-specific models that I designed for the image classification tasks, where I adopts various methods from metric learning to data augmentations to solve long-tail distribution problem (the main reason is, I think, that data to train the big model are curated as multi-modal style, in contrast, the task-specific model is trained only with classification label and image pairs).

Motivated from the experience, I have been in charge of investigating large-scale models, especially LLMs at SAPEON. In addition, as a member of transformer team, I designed the frontend for transformer acceleration tool ([P4]) and evaluated LLMs after quantization. Recently, for the purpose of a research project mentioned above (AI Simulation), I mostly have kept track of diffusion model-based architectures.

Lightweight and efficient model *May. 2017 -*

The ultimate goal of spiking neural network research team that I engaged in throughout graduate school is to develop energy-efficient neural network from algorithms to hardware design. With the development of lightweight methods, we naturally had focused on the quantization and its implementation with emerging analog memory devices. However, as these approaches lead to severe information loss, I conducted research on co-optimization between algorithms and implementations.

In this area, there are two approaches: one is to address issues derived from efficient hardware modeling and the other one from lightweight algorithm. At POSTECH, I conducted research on the former ([J1], [J2]). As the analog memory device is unstable than digital memory device, training a model with general algorithms leads to severe performance degradation, so it is required to modify these algorithms. With the modified algorithm which requires additional hyper-parameters, my research topics aimed to propose to optimize robust hyper-parameters. At SAEPON, I have developed quantization toolkit ([P4]) which aims to find optimal algorithms to lightweight models for NPUs.

In addition to develop quantization algorithms for SAPEON's NPUs, I have conducted research on diverse topics related to lightweight models ([T2]); scheduling with graph neural network based compiler in system level optimization like AutoTVM and generating synthetic data for calibration in conjunction with deep generative models ([P3],[C7],[C8]]).

Research Experiences

Hanyang Univ. *Jun. 2024 - Feb. 2025*

Motivated from conceptual experiments, I have proposed and led a research project with Hanyang univ. (Prof. Jungwook Choi) The final goal of the project is to propose optimization methods for text-guided diffusion models for pre-trained visual perception discriminative models. One of applications is to generate synthetic data to lightweight pre-trained discriminative models.

SRFC *Sep. 2020 - May. 2021*

This project founded by Samsung Research Funding and Incubation Center for Future Technology (SRFC) aims at solving a wide range of problems in analog circuit-based neural network with emerging resistive memory devices. Collaborating with IBM, our research team with Prof.Seyoung Kim as a principal investigator has conducted research on the following issues: 1) developing robust training or optimization algorithms, 2) co-optimizing SW and HW system, and 3) proposing an innovative simulator design methodology for designing edge devices.

Samsung Beyond Limit Project *Nov. 2018 - Feb.2020*

I had spearheaded this project with the object of developing analog neural network system with emerging memory device. From designing algorithms to fabrication, our research team collaborated with another research team which comes from a material science background. As a team leader of two research teams, I had guided the research direction under the supervision of Prof. Kiyoung Choi and Prof. Cheol Seong Hwang.

KIST Open Research Program *Mar. 2018 - Dec. 2018*

This project funded by Korea Institute of Science and Technology (KIST) mainly focused on designing a biologically-plausible learning algorithm for the spiking neural networks. Our research team covered various topics in spiking neural networks: Spiking-Time-Dependent Plasticity (STDP), equilibrium propagation, conversion to spiking neural networks, and temporal coding.

Working or Teaching Experiences

NPU/GPU Performance Engineer at Rebellions (M&A SAPEON Inc.) *Dec. 2024 - Jul. 2025*
As SAPEON Korea has merged with Rebelllions, I have engaged in Production Innovation team to evaluate the performance of NPU/GPU for LLM serving frameworks and develop applications for NPU. Along with it, I have conducted research on synthetic data generation using text-guided diffusion models, which is now under review ([P3], [C7], [C8]).

Research and Software Engineer at SAPEON Inc. *May. 2022 - Dec. 2024*
I has developed a quantization tool ([P4]) which is based on PyTorch while implementing kernels that emulate processing units and number formats of NPUs using C++ and CUDA. In addition to that, I have been in charge of developing LLM acceleration tool for NPUs, focusing on quantization and kernel optimization. To deal with quantization for data-free scenario, I have conducted research on the topic using text-guided diffusion models ([P3], [C7], [C8]).

Internship at NAVER Clova *July. 2021 - Jan. 2022*
Engaged in NAVER Clova OCR team as an internship, I participated the project that aims at building a neural network for classifying documents under the supervision of Bado Lee. In this project, I mainly focused on image retrieval with metric learning to deal with issues on open-set classification and domain generalization. In addition, with the project, I participated in the big-model project for multi-task learning.

Research Associate at POSTECH *Sept. 2020 - May. 2021*
Participating as a research associate in the department of Material Science and Engineering of Pohang University of Science and Technology, I conducted research on optimization theory and application for analog circuit-based neural networks using resistive memory devices, targeting for on-device learning. One of main topics is co-optimizing the hardware and software system. Especially, I analyzed the on-device learning algorithm based on the optimization theory ([J1,J2]). In addition, I have continued to develop the simulator which I started when I was engaged in ISRC ([A1, P2]).

Research Assistant at SNU/ISRC *May. 2017 - Aug. 2020*
I started research assistant engaged in the spiking neural networks research team at Design Automation Lab supervised by Prof. Kiyoung Choi. During the master course at SNU, my research topics were spread from quantized neural networks and its implementations ([C1, C4, C5, C6]), bio-inspired neural networks such as spiking neural networks and curriculum learning ([C2, C3]) , and distributed learning. One common thing of these topics is to design lightweight neural networks and co-optimize hardware systems. I conducted independent research on developing simulator for analog circuit-based neural networks with emerging memory devices. The simulator is targeting for Electronic Design Automation (EDA) of analog system using the concept of deep learning ([A1, P2]).

Internships at SK Hynix *Jun. 2015 - Aug. 2015*
I was a member of the analog circuit team and completed a project which reduces the time for setting-up gate threshold voltage in nano-level devices, resulting in the reduction of the delay time from 1ms to 1ns.

Preprints/Publications/Patents

Preprints

[A1] **Chaeun Lee** and Seyoung Kim, "SEMULATOR: Emulating the Dynamics of Crossbar Array-based Analog Neural System with Regression Neural Network," arXiv, 2021.

Publications

[J2] **Chaeun Lee**, Kyungmi Noh, Wonjae Ji, Tayfun Gokmen and Seyoung Kim, "Imapct of Asymmetric Weight Update on Neural Networks Training with Tiki-taka Algorithm," Frontiers in Neuroscience, 2022 (IF=4.7).

[J1] Hyunjung Kwak, Chuljun Lee, **Chaeun Lee**, Kyungmi Noh and Seyoung Kim, "Experimental Measurement of Ungated Channel Region Conductance in a Multi-terminal, Metal Oxide-based ECRAM", Semiconductor Science and Technology, 2021 (IF=2.36).

[C8] Jiwoong Park*, **Chaeun Lee***, Yongseok Choi, Sein Park, Deokki Hong, and Jungwook Choi, "Enhancing Generalization in Data-free Quantization via Mixup-class Prompting," ICCV BiVision Workshop (ICCVW), 2025 (*=equal contribution).

[C7] Jiwoong Park, **Chaeun Lee**, Yongseok Choi, and Jungwook Choi, "Data-scarce quantization using Stable diffusion," Autumn Annual Conference of IEIE, 2024 (domestic).

[C6] **Chaeun Lee**, Jaehyun Kim, and Kiyoung Choi, "An RRAM-based Analog Neuron Design for the Weighted Spiking Neural Network," International SoC Design Conference (ISOCC), 2019.

[C5] **Chaeun Lee**, Jaehyun Kim, Jihun Kim, Jaehyun Kim, and Kiyoung Choi, "Fast Simulation Method for Analog Deep Binarized Neural Networks," International SoC Design Conference (ISOCC), 2019.

[C4] Jaehyun Kim*, **Chaeun Lee***, Jihun Kim, Yumin Kim, Cheol Seong Hwang and Kiyoung Choi, "VCAM: Variation Compensation Technique through Activation Matching," International Symposium on Low Power Electronics and Design (ISLPED), 2019 (*=equal contribution, preliminary version appeared in Design Automation Conference Work-In-Progress (DAC WIP) 2019).

[C3] Jaehyun Kim, **Chaeun Lee**, and Kiyoung Choi, "Deep Neural Network Training with Random Search," SoC Conference, 2019 (domestic).

[C2] Seonghyun Jeong, **Chaeun Lee**, Jaehyun Kim, and Kiyoung Choi, "A Biologically Plausible Deep Learning Model with DBN and A Shallow Classifier," Institute of Semiconductor Engineers Conference, 2018 (domestic).

[C1] Jaehyun Kim, **Chaeun Lee**, and Kiyoung Choi, "Energy Efficient Analog Synapse/Neuron Circuit for Binarized Neural Networks," International SoC Design Conference (ISOCC), 2018.

Patents

[P4] **Chaeun LEE**, Deokki HONH, METHOD AND DEVICE FOR MODULARIZED COMPRESSION OF MACHINE LEARNING MODEL, Under examination

[P3] **Chaeun LEE**, METHOD AND DEVICE FOR GENERATING CALIBRATION DATASET CONSIDERING TRAINING DOMAIN OF NEURAL NETWORK MODEL AND FOR OPTIMIZING NEURAL NETWORK MODEL USING THE SAME, Under examination

[P2] **Chaeun LEE**, APPARATUS FOR CALCULATING EQUATIONS OF PROCESSING ELEMENTS USING NEURAL NETWORK AND METHOD FOR CONTROLLING THE SAME, Under examination

[P1] Ki Young CHOI, Jae Hyun KIM, **Chae Un LEE**, Joonyeon CHANG, Joon Young KWAK, Jaewook

KIM, METHOD FOR COMPENSATING FOR PROCESS VARIATION BY MEANS OF ACTIVATION VALUE ADJUSTMENT IN ANALOG BINARIZED NEURAL NETWORK CIRCUIT, AND SYSTEM THEREFOR, US Patent Application 20210089893, 2020.

Talks/Awards

Talks

[T4] **Chaeun Lee**, "Towards Efficient AI Agents: Models, Quantization, and Tools", Invited talk at Supergate Inc., 2025. [PPT]

[T3] **Chaeun Lee**, "Synthetic data generation with text-to-image diffusion models", Invited Talk at Hanyang univ., 2025. [PPT]

[T2] **Chaeun Lee**, "Recent advances in quantization for deep learning models from algorithms to system level ", **Main Tutorial** at International Conference on Electronics, Information, and Communication (ICEIC), 2024. [Official][PPT]

[T1] **Chaeun Lee**, "Statistical Modeling-based Simulators for Analog Neural Networks", **Main Tutorial** at International SoC Design Conference (ISOCC), 2021. [Official][PPT/Slide notes]

Awards

[A1] Jaehyun Kim, **Chaeun Lee**, and Kiyoung Choi, "Energy Efficient Analog Synapse/Neuron Circuit for Binarized Neural Networks," **Best Paper Award granted by SK Hynix** at International SoC Design Conference (ISOCC), 2018.

Skills

Language

English (TOEFL iBT 94), Korean (Native)

Programming Skills

Python, Machine Learning Frameworks (PyTorch, Tensorflow, ONNX), C/C++, CUDA, Linux

Academic Background

Mathematics, Electronics (circuit & physics), Computer science, Optimization, Neuroscience