Donghyun Son

happydh1@snu.ac.kr / Google Scholar / Personal Website

Education

Seoul National University, *B.S. in Computer Science* **March 2018 — Feb 2026 (expected)**

Three-year leave of absence for mandatory military service, served as an industrial technical personnel

- **GPA**: 4.04/4.3 (4.11/4.3 in major), expected to graduate ***Summa Cum Laude***

University of Texas at Austin, *Exchange Student, Computer Science* **Sep 2025 — Dec 2025**

Publications

Nalar*: A Serving Framework for Agent Workflows*
Benedict Marco Laju, **Donghyun Son**, Saurabh Agarwal, Nitin Kedia, Myungjin Lee, Jayanth Srinivasa, Aditya Akella
Under Review at **OSDI 2026**

NSNQuant: A Double Normalization Approach for Calibration-Free Low-Bit Vector Quantization of KV Cache
Donghyun Son, Euntae Choi, Sungjoo Yoo
Advances in Neural Information Processing Systems (**NeurIPS 2025**)

In-Context Learning with Noisy Labels
Junyong Kang, **Donghyun Son**, Hwanjun Song, Buru Chang
Preprint

Gradient Estimation for Unseen Domain Risk Minimization with Pre-Trained Models
Byounggyu Lew†, **Donghyun Son**†, Buru Chang
Workshop and Challenges for Out-of-Distribution Generalization in Computer Vision @ ICCV 2023 **(OOD-CV@ICCV 2023)**

Looking to Personalize Gaze Estimation Using Transformers
Seung Hoon Choi, **Donghyun Son**, Yunjong Ha, Seonghun Hong, Taejung Park
Journal of Computing Science and Engineering **(JCSE)**, Vol. 17, No. 2, pp.41-50

Reliable Decision from Multiple Subtasks through Threshold Optimization: Content Moderation in the Wild
Donghyun Son†, Byounggyu Lew†, Kwanghee Choi†, Yongsu Baek, Seungwoo Choi, Beomjun Shin, Sungjoo Ha, Buru Chang
ACM International Conference on Web Search and Data Mining **(WSDM 2023, Oral)**

Experience

Undergraduate Researcher **August 2025 — Present**
UTNS Research Group @ UT Austin *Austin, Texas*
Supervised by Prof. Aditya Akella

- Designing an agentic system where the central controller dynamically manages resource allocation, load balancing, and scheduling (submitted to OSDI26)

Undergraduate Research Intern **September 2024 — August 2025**
CMALab @ SNU *Seoul, South Korea*
Supervised by Prof. Sungjoo Yoo

- Investigated a calibration-free vector quantization method for KV cache, incorporating a novel normalization algorithm and efficient CUDA kernels for 1–2 bit inference (NeurIPS25)

Machine Learning Engineer, Moderation ML Team **March 2022 — July 2023**
Hyperconnect (Acquired by Match Group) *Seoul, South Korea*
Advised by Dr. Buru Chang

- Built an ML-based system for moderating images and audio across Match Group brands (e.g., Tinder, Hinge)
- Devised a threshold optimization algorithm based on GD to optimize the tradeoff between reliability and cost (WSDM23, Oral)
- Proposed a training algorithm for ML models to generalize effectively across different services and domains (OOD-CV@ICCV23)

Research Engineer, R&D Team **August 2020 — March 2022**
VisualCamp *Seoul, South Korea*

- Designed and implemented modern C++ based multi-threaded ML inference pipeline for gaze estimation SDK
- Built a robust training procedure to train an appearance-based gaze estimation model and improved accuracy by 30%
- Proposed a transformer-based calibration algorithm that predicts user-specific latent vectors using users' samples

Patents

***Apparatus and Method for Setting Criteria on Data Classification* (US20240281494A1)**
Yong Su Baek, **Dong Hyun SON**, Beom Jun Shin, Byoung Gyu Lew, Bu Ru CHANG, Kwang Hee CHOI, Seung Woo Choi, Sung Joo Ha

***Apparatus for Domain Generalization of Machine Learning Models, Methods and Computer Readable Recording Mediums Therefor* (US20240087294A1)**
Bu Ru CHANG, Byoung Gyu Lew, **Dong Hyun SON**

Honors & Awards

Programming Contests

- **2nd prize**, at Union of Clubs Programming Contest (UCPC) 2025
- **5th prize**, at Samsung Collegiate Programming Contest (SCPC) 2021
- **5th place**, at ICPC NERC Huawei Challenge 2020: Cloud Scheduling Challenge
- **1st place**, at Seoul National University Programming Contest (SNUPC) 2019, div. 2

Scholarships

- **Samsung Software Membership**, from Samsung Research (November 2021 — Present)
- **Korea-U.S. STEM Student Exchange Scholarship** ($9000), from Ministry of Trade, Industry and Energy (2025.09)
- **Full Tuition Academic Scholarship** (merit-based, $2300), from Seoul National University (2020-1, 2023-2, 2024-1, 2024-2)

Others

- **4th place**, at CXR-LT Challenge Task 1 (MICCAI 2024)
- **Bronze Prize**, at 42nd Undergraduate Mathematics Contest div.1 (for mathematical majors)
- **1st place**, at deep learning model acceleration challenge in HPC class (number of participants: 130+)
- **Microsoft Azure Champ Prize : Hack For Good**, at TartanHacks 2021 in CMU
- **Best Paper Award**, at Human & Cognitive Language Technology (HCLT) 2020

Presentations

- **KV Cache Compression for Long Context Inference**, Weekly Seminar @ Deepest S16 (January 2025)
- **Efficient Algorithms for LLM Inference**, Weekly Seminar @ Deepest S15 (July 2024)

Projects

RAG-based Research Assistant **Aug 2023 — Dec 2023**
In collaboration with SoftlyAI

- Developed a RAG-based research assistant that helps users understand papers in an interactive way
- Fine-tuned LLMs to align with human preferences and built a Milvus-based retrieval server

Leveraging in-context learning ability of LLMs for shallow fusion [github] **Oct 2023 — Dec 2023**

- Improved automatic speech recognition (ASR) by applying shallow fusion with an LLM conditioned on few-shot (ASR output, ground truth) correction examples.

Additional Information

Interests: Table Tennis, Math Puzzles, Mind Sports (4 Dan in Go)
Technical Skills: C++, Python, Pytorch, Tensorflow, Jax, CUDA, OpenCL, MPI, OpenMP
Problem Solving: Codeforces (handle: **diordhd**, rating: 2200+) / BOJ (handle: **dhdroid**)
TOEFL iBT: 108 (R29/L29/S22/W28)