# Donghyun Son

happydh1@snu.ac.kr / Google Scholar

## EDUCATION

**Seoul National University**, *B.S. in Computer Science*　　　　　　　　　　**March 2018 — Present**

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

- **GPA**: 4.05/4.3 (4.11/4.3 in major)
- **Relevant Coursework**: Scalable High-Performance Computing (Graduate Course); Natural Language Processing (Graduate Course); Computer Vision; Mathematical Statistics; Linear Algebra;

## EXPERIENCE

**Undergraduate Research Intern**　　　　　　　　　　**September 2024 — Present**

*CMALab @ SNU*　　　　　　　　　　*Seoul, South Korea*

*Supervised by prof. Sungjoo Yoo*

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

**Machine Learning Engineer, Moderation ML Team**　　　　　　　　　　**March 2022 — July 2023**

*Hyperconnect (of Match Group)*　　　　　　　　　　*Seoul, South Korea*

- Developed an ML-based system for moderating images and audio across Match Group brands (e.g., Tinder, Hinge)
- Designed an efficient human-in-the-loop system for content moderation, managing AI flywheel
- 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
- Accelerated a calibration process (personalizing gaze estimation) by 10x, by substituting naive optimization algorithms with L-BFGS and improving outlier removal process.
- 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

## PUBLICATIONS

*NSNQuant: A Double Normalization Approach for Calibration-Free Low-Bit Vector Quantization of KV Cache*

**Donghyun Son**, Euntae Choi, Sungjoo Yoo

arXiv Preprint, *Under Review*

*In-Context Learning with Noisy Labels*

Junyong Kang, **Donghyun Son**, Hwanjun Song, Buru Chang

arXiv 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 @ ICCV2023　**(OOD-CV@ICCV23)**

*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　**(WSDM23, Oral)**

## 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)

## Honors & Awards

- **Samsung Software Membership**, from Samsung Research (November 2021 — Present)
- **Bronze Medal**, at 42nd Undergraduate Mathematics Contest div.1 (for mathematical majors, by Korean Mathematical Society)
- **Full Tuition Academic Scholarship** (merit-based, $2300), from Seoul National University (2020-1, 2023-2, 2024-1)
- **5$^{th}$ award**, at Samsung Collegiate Programming Contest (SCPC) 2021
- **Microsoft Azure Champ Prize : Hack For Social Good**, at TartanHacks 2021 in CMU
- **Best Paper Award**, at HCLT 2020
- **5$^{th}$ place**, at ICPC NERC Huawei Challenge 2020: Cloud Scheduling Challenge
- **1$^{st}$ place**, at Seoul National University Programming Contest (SNUPC) 2019, div. 2

## Additional Information

**Interests**: Basketball, Swimming, Math Puzzles, Mind Sports (4 Dan in Go)
**English Proficiency**: scored 108 in TOEFL iBT (R: 29, L:29, S: 22, W: 28)
**Technical Skills**: C++, Python, Pytorch, Tensorflow, Jax, CUDA, OpenCL, MPI, OpenMP
**Problem Solving**: **diordhd** in codeforces (2200+) / **dhdroid** in BOJ