Wookje Han

347-971-1195 | wookje.han@columbia.edu | LinkedIn

Education

Columbia University — New York, NY
MS in Computer Science — *Aug 2023 - Dec 2024 (Expected)*
Seoul National University (SNU) — Seoul, Korea
BS in Computer Science and Engineering, Graduated Summa cum laude — *Mar 2017 - Aug 2023*

Technical Skills

Languages: Python, C/C++, Java, OCaml, R, Rust, Kotlin
Frameworks, Libraries: PyTorch, Transformers, DeepSpeed, NumPy, Matplotlib, Hydra, Weights & Biases, pybind, OpenMP, MPI, CUDA
Developer Tools: Git, Docker, AWS, GCP, Latex

Experience

Undergraduate Research Assistant @ SNU — Seoul, Korea
Software Platform Lab — *Dec 2021 - Feb 2023*
- Distributed memory load of large language models by parallelizing across multiple GPUs with MPI and PyTorch
- Integrated Hydra to project code, aiding team members to manage configurations efficiently
- Automated parallelization correctness validation process by constructing a unit test, enabling efficient validation
- Designed a task description generation algorithm, enhancing large language model's few-shot performance by 19%

Language and Data Intelligence Lab — *Jun 2021 - Dec 2021*
- Built a 4x more efficient algorithm for continual knowledge update in language models with Transformers
- Established an environment with docker, enabling team members to conduct experiments with the same settings
- Implemented a method for Multi-Hop QA with PyTorch, leading to 5% performance gain

Programming Research Lab — *Sep 2019 - Dec 2019*
- Developed a static analyzer with OCaml to validate program safety by type checking

Software Engineer Intern @ Samsung Electronics — Suwon, Korea
MX Mobile AI Server Team — *Jul 2019 - Aug 2019*
- Assisted in designing database table and implementing backend process for AI assistant (Bixby), reducing data storage redundancy and response latency leveraging AWS DynamoDB

Projects

Shot Selection Optimization | *Python, PyTorch, Git, Weights & Biases, Linux* — *Sep 2022 – Dec 2022*
- Proposed selecting few shots via RL, improving language model's few-shot performance up to 12.2% on average

LLVM Compiler Optimization | *C++, LLVM Compiler, Git, Docker, Linux* — *Mar 2022 – Jul 2022*
- Led a team of 4, devising optimizations for LLVM compiler and reduced 42% of cost for a given virtual machine

Dataset Corruption Detection | *Python, PyTorch, Git, CrypTen, Linux* — *Mar 2022 – Jul 2022*
- Experimented impact of corrupted dataset in multi-party deep learning, underscoring the need for detection
- Applied Zero Knowledge Proof to detect dataset corruption in multi-party deep learning while securing privacy

FastDCGAN | *C, MPI, CUDA, Linux* — *Sep 2021 – Dec 2021*
- Accelerated DCGAN 200X faster by exploiting multiple GPU nodes using CUDA and MPI

ArchPresser | *Python, NumPy, pybind, Git, Linux* — *Sep 2021 – Dec 2021*
- Constructed a program to generate a panoramic dental x-ray image from 3D CT scans using Python and C++
- Analyzed bottleneck and used pybind to integrate C++ code with Python, making it 20x faster

EZOrder | *Kotlin, Java, Android Studio, Git, Linux* — *Sep 2019 – Dec 2019*
- Developed frontend process of Android application to handle customer orders with Kotlin
- Proposed an application's expected time prediction algorithm using queuing theory

Publications

- **Wookje Han**, TaeHyun Lee, Hyeonmin Ha, Byung-Gon Chun. "A Survey on Memory Optimization Techniques and Frameworks for Training Large Language Models." *KSC (2022)*
- Kyungjae Lee, **Wookje Han**, Seung-Won Hwang, Hwaran Lee, Joonsuk Park and Sang-Woo Lee. "Plug-and-Play Adaptation for Continuously-updated QA." *Findings of ACL (2022)*
- (*Under Review*) Hyeonmin Ha, JIHYE LEE, **Wookje Han**, Byung-Gon Chun. Improved the performance of large language models that are deployed as a service (*i.e* GPT4, GPT-3.5) by generating an automatic prompt
- (*Under Review*) **Wookje Han**, Jinsol Park, Kyungjae Lee. Designed a unified system that enables large language models to handle any type of information-seeking question by exploiting presuppositions