Dr. Hayeon Lee

https://hayeonlee.github.io/
yeonhi926@gmail.com
Google Scholar GitHub LinkedIn

Research Interests

My research focuses on developing efficient large-scale AI models for real-world applications. I specialize in model optimization techniques, including AutoML, Neural Architecture Search (NAS), meta-learning, hyperparameter optimization, and model compression through Knowledge Distillation (KD). Recently, I have been particularly interested in designing, analyzing, and evaluating large language models (LLMs) for efficient long-context generation and optimizing Retrieval-Augmented Generation (RAG) techniques, enabling practical deployment in real products.

Keywords: Efficient LLMs; RAG; AutoML; NAS; KD

Work Experience

Research Scientist Oct. 2024 - Present
Meta GenAI, Menlo Park, CA, United States
- Topic: Optimizing Retrieval-Augmented Generation (RAG) techniques, with a focus on improving memory and latency efficiency.

Postdoctoral Researcher Sep. 2023 - Sep. 2024
Meta FAIR Labs, Menlo Park, CA, United States
worked with **Dr. Yuandong Tian**
- Topic: Developed a key-value (KV) cache compression approach with retrieval heads to enable memory-efficient long-context generation using LLMs.
- Paper: MLSys 2025 Submitted [P3]

Research Scientist Intern Aug. 2022 - Dec. 2022
Meta AI, Seattle, WA, United States
worked with **Dr. Alex Min**
- Topic: Developed and analyzed KD techniques for improving the quality and compressing language models.
- Paper: ACL Findings 2023 [C7], EMNLP Findings 2023 [C8]

Research Scientist Intern Jul. 2015 - Aug. 2015
National AI Research Institute, Daejeon, South Korea

Software Engineer Intern Jan. 2015 - Feb. 2015
Samsung Electronics, Suwon, South Korea

Honors

Best Ph.D. Dissertation Award, College of Engineering, KAIST, 2024

Google Travel Grant for ICLR, Google, 2023

Spotlight Presentation, ICLR, 2023

Keynote Speaker, AutoML Conference, 2023

Google AI/CS/EE Rising Stars 2023, Google Explore Computer Science Research, 2023

Google Ph.D. Fellowship, Research Area: Machine Learning, Google Research, 2022

Outstanding Reviewer, NeurIPS, 2022

Spotlight Presentation, ICLR, 2022

Google AI/CS/EE Rising Stars 2022, Google Explore Computer Science Research, 2022

Spotlight Presentation, NeurIPS, 2021

Best Presentation Award, Korea Agency for Defense Development Workshop, 2021

Naver Ph.D. Fellowship, NAVER Corp., 2020

Outstanding Reviewer, ICML, 2020

Kyunghyun Cho Travel Grant for ICLR, KAIST, 2020

Oral Presentation, ICLR, 2020

Publications

* denotes equal contribution

Preprints

[P3] KV Cache Compression with Retrieval Heads for Efficient Long Context Factuality
Hayeon Lee and Yuandong Tian
Arxiv. 2024

[P2] Diffusion-based Neural Network Weights Generation
Bedionita Soro, Bruno Andreis, **Hayeon Lee**, Song Chong, Frank Hutter, Sung Ju Hwang
Arxiv. 2024

[P1] SuperNet in Neural Architecture Search: A Taxonomic Survey
Stephen Cha, Taehyeon Kim, **Hayeon Lee**, Se-Young Yun
Arxiv. 2022

Conferences

[C9] DiffusionNAG: Predictor-guided Neural Architecture Generation with Diffusion Models
Sohyun An*, **Hayeon Lee***, Sung Ju Hwang
International Conference on Learning Representations (ICLR) 2024

[C8] Co-training and Co-distillation for Quality Improvement and Compression of Language Models
Hayeon Lee, Jongpil Kim, Rui Hou, Davis Liang, Hongbo Zhang, Sung Ju Hwang, Alexander Min
Findings of Empirical Methods in Natural Language Processing (EMNLP) 2023

[C7] A Study on Knowledge Distillation from Weak Teacher for Scaling Up Pre-trained Language Models
Hayeon Lee, Rui Hou, Jongpil Kim, Davis Liang, Sung Ju Hwang and Alexander Min
Findings of Association for Computational Linguistics (ACL) 2023

[C6] Meta-Prediction Model for Distillation-aware NAS on Unseen Datasets
Hayeon Lee*, Sohyun An*, Sung Ju Hwang
International Conference on Learning Representations (ICLR) 2023
Spotlight Presentation, (notable-top-25%)

[C5] Online Hyperparameter Meta-Learning with Hypergradient Distillation
Hae Beom Lee, **Hayeon Lee**, Jaewoong Shin, Eunho Yang, Timothy Hospedales, Sung Ju Hwang
International Conference on Learning Representations (ICLR) 2022
Spotlight Presentation, (acceptance ratio = 176/3391 = 5.1%)

[C4] HELP: Hardware-Adaptive Efficient Latency Predictor for NAS via Meta-Learning
Hayeon Lee*, Sewoong Lee*, Chong Song, Sung Ju Hwang
Conference on Neural Information Processing Systems (NeurIPS) 2021
Spotlight Presentation, (acceptance ratio $< 3\%$)

[C3] Task-Adaptive Neural Network Search with Meta-Contrastive Learning
Wonyong Jeong*, **Hayeon Lee***, Gun Park*, Eunyoung Hyung, Jinheon Baek, Sung Ju Hwang
Conference on Neural Information Processing Systems (NeurIPS) 2021
Spotlight Presentation, (acceptance ratio $< 3\%$)

[C2] Rapid Neural Architecture Search by Learning to Generate Graphs from Datasets
Hayeon Lee*, Eunyoung Hyung*, Sung Ju Hwang
International Conference on Learning Representations (ICLR) 2021

[C1] Learning to Balance: Bayesian Meta-Learning for Imbalanced and Out-of-distribution Tasks
Hae Beom Lee*, **Hayeon Lee***, Donghyun Na*, Saehoon Kim, Minseop Park, eunho Yang, Sung Ju Hwang
International Conference on Learning Representations (ICLR) 2020
Oral Presentation, (acceptance ratio = 48/2594 = 1.9%)

Workshops

[W1] Lightweight Neural Architecture Search with Parameter Remapping and Knowledge Distillation
Hayeon Lee*, Sohyun An*, Minseon Kim, Sung Ju Hwang
First Conference on Automated Machine Learning (Late-Breaking Workshop) (AutoML) 2022

Keynote

"Transferable Neural Architecture Search with Diffusion Models for the Real World" — Sep. 2023
AutoML Conference (Germany)

Invited Talks

"Rapid Neural Architecture Search by Learning to Generate Graphs from Datasets" — Apr. 2021
Samsung Electronics DS DIT Center (South Korea)

"Rapid Neural Architecture Search by Learning to Generate Graphs from Datasets" — Oct. 2021
Agency for Defense Development (South Korea)

"Task-Adaptive Neural Network Search with Meta-Contrastive Learning" — Dec. 2021
NeurIPS Social: ML (remote)

"Task-Adaptive Neural Network Search with Meta-Contrastive Learning" — Apr. 2022
Hanbat National University (South Korea)

"Task-Adaptive Neural Network Search with Meta-Contrastive Learning" — May. 2022
KAIST Programming Language Research Group (South Korea)

"Task-Adaptive Neural Network Search with Meta-Contrastive Learning" — May. 2022
Electronic & Information Research Information Center (South Korea)

"HELP: Hardware-Adaptive Efficient Latency Prediction for NAS via Meta-Learning" — Dec. 2021
NeurIPS Social: ML (remote)

"HELP: Hardware-Adaptive Efficient Latency Prediction for NAS via Meta-Learning" — Apr. 2022
Hanbat National University (South Korea)

"HELP: Hardware-Adaptive Efficient Latency Prediction for NAS via Meta-Learning" — Jun. 2022
Ewha University (South Korea)

Academic Activities

Area Chair: AutoML 24

Online Experience Chair: AutoML 24

Conference Reviewer: NeurIPS 20-24, ICML 20-24, ICLR 21-25, CVPR 23-24, AAAI 21, ACML 20, ACL ARR. 22

Journal Reviewer: TMLR

Industrial Project

Human-Inspired Large-Scale Visual Recognition System — Feb. 2019 - Dec. 2022
Samsung Electronics (South Korea)

AutoML with Large-scale Hyperparameter Meta-Learning — Dec. 2022 - Aug. 2023
Google Inc. (South Korea)

Tech. Skills

Programming: Python, MATLAB
Machine Learning: PyTorch, Huggingface Transformers

Mentoring

Sohyun An, Ph.D. Student at UCLA — Apr. 2022 - Aug. 2024
- Topic: Neural Architecture Search, Diffusion Models
- Paper: AutoML 2022 [W1], ICLR 2023 [C6], ICLR 2024 [C9]

Sewoong Lee, Ph.D. Student at KAIST — Feb. 2021 - Dec. 2021
- Topic: Neural Architecture Search
- Paper: NeurIPS 2021 [C4]

Eunyoung Hyung, AI Researcher at Samsung Research — Sep. 2019 - Jan. 2021
- Topic: Neural Architecture Search
- Paper: ICLR 2021 [C2], NeurIPS 2021 [C3]

Education

Ph.D. in School of Computing — Mar. 2018 - Aug. 2023
KAIST, Daejeon, South Korea
Advisor: **Prof. Sung Ju Hwang**
Dissertation Title: "Efficient and Generalizable Neural Architecture Search for the Real World"
- Best Ph.D. Dissertation Award from College of Engineering, KAIST, 2024
- Committee: **Prof. Sung Ju Hwang**, **Prof. Frank Hutter**, **Prof. Cho-Jui Hsieh**, **Prof. Eunho Yang**, **Prof. Se-Young Yun**

M.S. in School of Computing — Mar. 2016 - Feb. 2018
KAIST, Daejeon, South Korea

B.S. in Computer Science — Mar. 2012 - Feb. 2016
Sungkyunkwan University, Suwon, South Korea

Reference

Prof. Sung Ju Hwang, Endowed Chair Professor @ KAIST
Contact: sjhwang82@kaist.ac.kr

Dr. Yaundong Tian, Research Scientist Director at Meta FAIR
Contact: yuandong@meta.com

Dr. Alex Min, Research Scientist at Meta AI
Contact: alexmin@meta.com

Prof. Frank Hutter, Full Professor @ University of Freiburg
Contact: fh@cs.uni-freiburg.de

last update: September 2024