Gyeong-In Yu

CTO of FriendliAI

1 Gwanak-ro, Gwanak-gu Bldg. 138, Rm. 514; Seoul, 08826, Korea

+82-10-9511-9531 | gyeonginyu at gmail dot com | gyeongin.github.io | gyeongin | gyeonginyu

Research Interests

My research interest lies in the intersection of computer systems and machine learning, with a focus on systems for machine learning. During my Ph.D., I primarily worked on software systems for accelerating machine learning inference in the datacenter.

Education

Seoul National University — *Seoul, Korea*
Ph.D. in Computer Science and Engineering — *Mar. 2017 - Feb. 2023*
- Advisor: Prof. Byung-Gon Chun

Seoul National University — *Seoul, Korea*
B.S. in Computer Science & Engineering and B.A. in Economics — *Mar. 2012 - Feb. 2017*
- Computer Science & Engineering GPA: 4.19/4.3

Korea Science Academy (KSA) of KAIST — *Busan, Korea*
Mar. 2009 - Feb. 2012

Professional Experience

CTO of FriendliAI — *Seoul, Korea*
Mar. 2023 - present

Research Intern at Microsoft AI and Research — *Redmond, WA*
Mentor: Matteo Interlandi, Saeed Amizadeh — *Jun. 2018 - Sep. 2018*

Research Intern at Microsoft Research Asia — *Beijing, China*
Mentor: Ming Wu — *Jun. 2017 - Sep. 2017*

Publications

Google Scholar: `https://scholar.google.com/citations?user=RwhPHaEAAAAJ`

Conference Publications

1. **Gyeong-In Yu**, Saeed Amizadeh, Sehoon Kim, Artidoro Pagnoni, Ce Zhang, Byung-Gon Chun, Markus Weimer, Matteo Interlandi. WindTunnel: Towards Differentiable ML Pipelines Beyond a Single Model. *48th International Conference on Very Large Data Bases (VLDB 2022).* [paper]
2. **Gyeong-In Yu**, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, Byung-Gon Chun. Orca: A Distributed Serving System for Transformer-Based Generative Models. *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2022)*, July 2022. [paper]
3. Taebum Kim, Eunji Jeong, Geon-Woo Kim, Yunmo Koo, Sehoon Kim, Gyeong-In Yu, Byung-Gon Chun. Terra: Imperative-Symbolic Co-Execution of Imperative Deep Learning Programs. 35th Conference on Neural Information Processing Systems (NeurIPS 2021), December 2021. [paper]

4. Woosuk Kwon*, **Gyeong-In Yu***, Eunji Jeong, Byung-Gon Chun (*equal contribution). Nimble: Lightweight and Efficient GPU Task Scheduling for Deep Learning. *34th Conference on Neural Information Processing Systems (NeurIPS 2020) (Spotlight)*, December 2020. [paper]
5. Supun Nakandala, Karla Saur, **Gyeong-In Yu**, Konstantinos Karanasos, Carlo Curino, Markus Weimer, Matteo Interlandi. A Tensor Compiler Approach for One-size-fits-all ML Prediction Serving. *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2020)*, November 2020. [paper]
6. Woo-Yeon Lee, Yunseong Lee, Joo Seong Jeong, **Gyeong-In Yu**, Joo Yeon Kim, Ho Jin Park, Beomyeol Jeon, Wonwook Song, Gunhee Kim, Markus Weimer, Brian Cho, Byung-Gon Chun. Automating System Configuration of Distributed Machine Learning. *39th IEEE International Conference on Distributed Computing Systems (ICDCS 2019)*, July 2019. [paper]
7. Soojeong Kim, **Gyeong-In Yu**, Hojin Park, Sungwoo Cho, Eunji Jeong, Hyeonmin Ha, Sanha Lee, Joo Seong Jeong, Byung-Gon Chun. Parallax: Sparsity-aware Data Parallel Training of Deep Neural Networks. *14th European Conference on Computer Systems (EuroSys 2019)*, March 2019. [paper]
8. Eunji Jeong, Sungwoo Cho, **Gyeong-In Yu**, Joo Seong Jeong, Dongjin Shin, Byung-Gon Chun. JANUS: Fast and Flexible Deep Learning via Symbolic Graph Execution of Imperative Programs. *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2019)*, February 2019. [paper]
9. Eunji Jeong*, Joo Seong Jeong*, Soojeong Kim, **Gyeong-In Yu**, Byung-Gon Chun (*equal contribution). Improving the Expressiveness of Deep Learning Frameworks with Recursion. *13th European Conference on Computer Systems (EuroSys 2018)*, April 2018. [paper]

Other Publications

1. Supun Nakandala, **Gyeong-In Yu**, Markus Weimer, Matteo Interlandi. Compiling Classical ML Pipelines into Tensor Computations for One-size-fits-all Prediction Serving. *Systems for ML Workshop at 33rd Conference on Neural Information Processing Systems (NeurIPS)*, December 2019. [paper]
2. Ahnjae Shin, Dong-Jin Shin, Sungwoo Cho, Do Yoon Kim, Eunji Jeong, **Gyeong-In Yu**, Byung-Gon Chun. Stage-based Hyper-parameter Optimization for Deep Learning. *Systems for ML Workshop at 33rd Conference on Neural Information Processing Systems (NeurIPS)*, December 2019. [paper]
3. Eunji Jeong, Sungwoo Cho, **Gyeong-In Yu**, Joo Seong Jeong, Dong-Jin Shin, Taebum Kim, Byung-Gon Chun. Speculative Symbolic Graph Execution of Imperative Deep Learning Programs. *ACM SIGOPS Operating Systems Review (OSR)*, July 2019. [paper]
4. Eunji Jeong, Sungwoo Cho, **Gyeong-In Yu**, Joo Seong Jeong, Dongjin Shin, Byung-Gon Chun. Demonstration of JANUS: Fast and Flexible Deep Learning via Symbolic Graph Execution of Imperative Programs. *Demonstration at Second Conference on Machine Learning and Systems (MLSys)*, April 2019. [paper]
5. **Gyeong-In Yu**, Saeed Amizadeh, Byung-Gon Chun, Markus Weimer, Matteo Interlandi. Making Classical Machine Learning Pipelines Differentiable: A Neural Translation Approach. *Systems for ML Workshop at 32nd Conference on Neural Information Processing Systems (NeurIPS)*, December 2018. [paper]
6. Soojeong Kim, Eunji Jeong, Joo Seong Jeong, **Gyeong-In Yu**, Hojin Park, Byung-Gon Chun. Auto-Parallelizing Deep Learning for Multi-machine, Multi-GPU Environments. *Workshop on AI Systems at 26th ACM Symposium on Operating Systems Principles (SOSP)*, October 2017.
7. Byung-Gon Chun, Brian Cho, Beomyeol Jeon, Joo Seong Jeong, Gunhee Kim, Joo Yeon Kim, Woo-Yeon Lee, Yun Seong Lee, Markus Weimer, **Gyeong-In Yu**. Dolphin: Runtime Optimization for Distributed Machine Learning. *ML Systems Workshop at 33rd international conference on machine learning (ICML)*, June 2016.

Patents

1. **Gyeongin Yu**, Geon-Woo Kim, Joo Seong Jeong, Soojeong Kim, Byung-Gon Chun. Selective batching for inference system for transformer-based generation tasks. US Patent 11,514,370, 2022.

2. **Gyeongin Yu**, Geon-Woo Kim, Joo Seong Jeong, Soojeong Kim, Byung-Gon Chun. Dynamic batching for inference system for transformer-based generation tasks. US Patent 11,442,775, 2022.

Honors, Awards, Grants

2023 **Best Ph.D. Dissertation Award**
SNU CSE

2021 **Youlchon AI Star**
Nongshim Youlchon Foundation and SNU AI Institute

Undergrad **National Scholarship for Science and Engineering**
Korea Student Aid Foundation

Teaching Experience

Fall 2019 **Theory and Lab of IoT, AI, and Big Data** *Teaching Assistant*
Seoul National University

Fall 2018 **Big Data Analytics and Deep Learning Systems** *Teaching Assistant*
Seoul National University

Spring 2017 **Operating Systems** *Teaching Assistant*
Seoul National University