Byung Soo Jeon

Distributed Inference @ NVIDIA | CS PhD @ CMU | U.S. Lawful Permanent Resident (LPR)

412-628-3003 | soojeonml@gmail.com | madfunmaker.github.io

Profile

I am a research-minded engineer specializing in high-performance distributed ML systems for LLMs, with full-stack experience across both training and inference spanning ML compiler optimization, multi-GPU parallelism, operator fusion, and GPU kernel development.

Professional Experience

Sep 2024 - **Senior System SW Engineer,** Multi-GPU inference and performance tech lead for TensorRT compiler — *NVIDIA*

- Led design and development of multi-GPU support for TensorRT's compiler backend from scratch
- Optimized context/tensor parallelism end-to-end, from graph-level op fusion to custom CUDA kernels

Jul - Sep 2024 **Staff ML System Engineer,** Optimized batching for a distributed LLM inference engine — *OctoAI (now NVIDIA)*

2017 - 2024 **Research Assistant,** Thesis: Automated & Portable Machine Learning System — *CMU*

- Built ML compiler/distributed runtime that delivers portable, optimized performance for Transformers — up to 1.6x speedup (training) and 1.4x (inference) over baseline ML frameworks (see publications)
- Automated graph-level optimizations, reducing manual tuning effort from days to minutes

Summer 2023 **Research Intern,** Benchmarked parallelisms for LLM inference and its implication on HW/SW co-design — *Google*

Summer 2020 **Applied Scientist Intern,** Researched on efficient meta-reinforcement learning and active exploration — *Amazon*

Summer 2019 **Applied Scientist Intern,** Developed a simulator / RL agent that solves online binpacking problem — *Amazon*

Jan - Jun 2017 **Research Intern,** Developed end-to-end multi-modal neural network for music emotion recognition — *Naver*

2015 - 2016 **Researcher,** Developed a distributed system / algorithm for billion-scale tensor algebra — *KAIST / SNU*

2013 **Co-founder & SW Engineer,** Developed client and server for multiplayer racing mobile game — *Funpresso, Inc*

Education

CMU (Carnegie Mellon University) — *Pittsburgh, PA*

Ph.D. in Computer Science — *May 2024*

- Thesis : Automated and Portable Machine Learning System | Committee: Tianqi Chen (Co-chair), Zhihao Jia (Co-chair), Greg Ganger, Luis Ceze

KAIST (Korea Advanced Institute of Science and Technology) — *Daejeon, Korea*

B.S. in Computer Science (Summa Cum Laude) — *Aug 2015*

Publications

Automated and Portable ML System

Cache Parallelism: Comparative Analysis of Parallelisms in Distributed LLM Inference for Long Sequence — *Thesis Chapter*

Byungsoo Jeon, Tianqi Chen, Zhihao Jia

GraphPipe: Improving the Performance and Scalability of DNN Training with Graph Pipeline Parallelism — *ASPLOS 2025*

Byungsoo Jeon*, Mengdi Wu*, Sunghyun Kim*, Shiyi Cao*, Sunghyun Park, Neeraj Aggarwal, Colin Unger, Daiyaan Arfeen, Peiyuan Liao, Xupeng Miao, Mohammad Alizadeh, Gregory R. Ganger, Tianqi Chen, Zhihao Jia

Collage: Seamless Integration of Deep Learning Backends with Automatic Placement — *PACT 2022*

Byungsoo Jeon*, Sunghyun Park*, Peiyuan Liao, Sheng Xu, Tianqi Chen, Zhihao Jia

- Integrated to Apache TVM Open-source Project (v0.9.0) | Presented in GTC 2022

SRTuner: Effective Compiler Optimization Customization By Exposing Synergistic Relations — *CGO 2022*

Sunghyun Park, Salar Latifi, Yongjun Park, Armand Behroozi, **Byungsoo Jeon**, Scott Mahlke

Applied ML / RL

FactoredRL: Leveraging factored graphs for deep reinforcement learning — *NeurIPS 2020*

Bharathan Balaji*, Petros Christodoulou*, Xiaoyu Lu*, **Byungsoo Jeon**, Jordan Bell-Masterson — *DeepRL Workshop*

OBP-RL: Exploring Deep Reinforcement Learning Methods for Online Binpacking Problem — *AMLC 2020*

Byungsoo Jeon, Bharathan Balaji, Saurabh Gupta, Chun Ye — *Amazon ML Conf*

Dropout Prediction over Weeks in MOOCs by Learning Representations of Clicks and Videos *AAAI 2020*
Byungsoo Jeon*, Namyong Park* *AI4Edu Workshop*

Dropout Prediction over Weeks in MOOCs via Interpretable Multi-Layer Representation Learning *AAAI 2020*
Byungsoo Jeon*, Namyong Park*, Seojin Bang* *AI4Edu Workshop*

Time-series Insights into the Process of Passing or Failing Online University Courses using Neural-Induced Interpretable Student States *EDM 2019*
Byungsoo Jeon, Eyal Shafran, Luke Breitfeller, Jason Levin, Carolyn P. Rose

Attentive Interaction Model: Modeling Changes in View in Argumentation *NAACL 2018*
Yohan Jo, Shivani Poddar, **Byungsoo Jeon**, Qinlan Shen, Carolyn P. Rose, Graham Neubig

Music Emotion Recognition via End-to-End Multimodal Neural Networks *RecSys 2017*
Byungsoo Jeon, Chanju Kim, Adrian Kim, Dongwon Kim, Jangyeon Park, Jungwoo Ha *Poster*

Distributed System / Algorithm for Tensor Algebra

BIGtensor: Mining Billion-Scale Tensor Made Easy *CIKM 2016*
Namyong Park*, **Byungsoo Jeon***, Jungwoo Lee, U Kang *Demo paper*

SCouT: Scalable Coupled Matrix-Tensor Factorization - Algorithm and Discoveries *ICDE 2016*
Byungsoo Jeon, Inah Jeon, Lee Sael, U Kang

TeGViz: Distributed Tera-Scale Graph Generation and Visualization *ICDM 2015*
ByungSoo Jeon, Inah Jeon, U Kang *Demo paper*

Teaching

Spring 2021	**Machine Learning Systems**, TA (Instructor: Tianqi Chen)	*CMU*
Spring 2020	**Deep Reinforcement Learning and Control**, TA (Instructor: Katerina Fragkiadaki)	*CMU*
Spring 2019	**Machine Learning (PhD)**, TA (Instructors: Leila Wehbe, Aaditya Ramdas)	*CMU*

Fellowship

2022	**Qualcomm Innovation Fellowship**, One of 19 winners in US ($100k for an year)	*Qualcomm*
2017 - 2021	**Kwanjeong Scholarship**, One of ~ 50 nationwide outstanding PhD students in STEM ($30k per year)	*KEF*

Skills

ML / Distributed System C++, Python | TensorRT, PyTorch, TVM, vLLM | CUDA, NCCL, SLURM, MPI