Byung Soo Jeon

SENIOR SYSTEM ENGINEER @ NVIDIA | CS PHD @ CMU | U.S. LAWFUL PERMANENT RESIDENT (LPR)

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

Passion

I am a research-minded engineer passionate about co-designing algorithm and system for efficient multimodal and large language models. I have experience building automated and portable distributed ML systems with a focus on parallelism, operator fusion, and graph optimizations.

Professional Experience

Sep 2024 -	**Senior System SW Engineer**, Building a compiler for distributed Transformer inference in TensorRT	*NVIDIA*
Jul - Sep 2024	**Staff ML System Engineer**, Optimized scheduling for a distributed LLM inference engine	*OctoAI (now NVIDIA)*
2017 - 2024	**Research Assistant**, Thesis: Automated & Portable Machine Learning System	*CMU*
Summer 2023	**Research Intern**, Investigated parallelisms for LLM inference and its implication on HW/SW co-design	*Google*
Summer 2020	**Applied Scientist Intern**, Research on efficient meta-reinforcement learning and active exploration	*Amazon*
Summer 2019	**Applied Scientist Intern**, Research on reinforcement learning for online combinatorial optimization	*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, and Jungwoo Ha *Poster*

Distributed System / Algorithm for Tensor Algebra

BIGtensor: Mining Billion-Scale Tensor Made Easy *CIKM 2016*

Namyong Park*, **Byungsoo Jeon***, Jungwoo Lee, and U Kang *Demo paper*

SCouT: Scalable Coupled Matrix-Tensor Factorization - Algorithm and Discoveries *ICDE 2016*

Byungsoo Jeon, Inah Jeon, Lee Sael, and U Kang

TeGViz: Distributed Tera-Scale Graph Generation and Visualization *ICDM 2015*

ByungSoo Jeon, Inah Jeon, and 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 | PyTorch, TensorRT, vLLM, TVM | CUDA, NCCL, SLURM, MPI, Docker