Heungsub Lee

Contact
heungsub@subl.ee

Web Sites
subl.ee, github.com/sublee, linkedin.com/in/sublee

Interests

- Software for software engineers or researchers
- Parallelism and distributed systems
- Cost optimization and management, especially for GPU-intensive systems

Skills

Programming Languages
Go, Python, TypeScript, JavaScript, Bash

Back-end Development
Linux, K8s, AWS, Terraform, ZeroMQ, NoSQL

ML Engineering
PyTorch, pipeline parallelism, NVIDIA Nsight Systems

Work Experience

Software Engineer
Global AI Platform Corporation, Sep 2023 – Present
Developing and maintaining LLM-based services such as Gistty, a Chrome extension summarizing Amazon reviews.

Software Engineering Manager
NAVER, Aug 2020 – Jul 2023
Supervised MLOps platforms by leading 25 software engineers to optimize the inference performance and productivity of HyperCLOVA, an LLM specializing in Korean culture.

Developed the second version of NSML, an ML research platform in CLOVA, to extend its capability for large-scale AI models on HPC infrastructures.

Software Engineer
Kakao Brain, Dec 2018 – Aug 2020
Developed a serverless training framework and a distributed hyperparameter search platform for an AutoML service.

Developed and published a pipeline parallelism library named torchgpipe in open source.

Game Server Engineer
NEXON, Mar 2011 – Dec 2018
Developed cloud-based distributed MMORPG servers for Durango using pub/sub communication over the spatial grid system. Achieved up to 70k concurrent users per game world.

Developed online racing game servers and matchmaking for KartRider Dash and KartRider Coin Rush.

Back-end Web Developer
nPine, Dec 2008 – Feb 2011
Developed web services selling stock images.

Front-end Web Developer
Lunant, Dec 2007 – Jan 2011
Designed and developed UI/UX for social media.

Open Source Experience

torchgpipe, Feb 2019 – Apr 2020
Implemented GPipe, a multi-GPU pipeline parallelism technique for training giant models, as a PyTorch library with optimization for CUDA, the autograd engine, and long skip connections. This project has become a part of PyTorch as Pipe APIs

Hangulize, Oct 2010 – Present
Invented a Hangul transcription algorithm and served as a web tool at zero cost.

Many professional Korean translators use this tool to translate undocumented proper nouns. Netflix refers to this tool in the Korean timed-text style guide.

TrueSkill, Jan 2012 – Dec 2015
Implemented TrueSkill™, the rating algorithm for Xbox Live, as a Python library. This project was introduced in PyData Berlin 2019.

Profiling, Aug 2014 – Nov 2017
Developed a Python profiler with an interactive TUI inspired by the Unity profiler. It was the 3rd daily trending repository in GitHub on Sep 22, 2014.

Contributions
- For PyTorch, fixed potential GPU memory violation (#27371); deprecated inconsistent API (#21006, #25985); discussed a counterintuitive behavior of the autograd engine (#18568).
- For ZeroMQ, discussed a PUB socket crash (#2942).
- For Flask, fixed a bug to generate a URL with a subdomain (#108).
- For jQuery 1.4.3, fixed a bug on content negotiation in Ajax requests.

Publications

- B. Kim et al., "What changes can large-scale language models bring? Intensive study on HyperCLOVA: Billions-scale Korean generative pretrained Transformers," arXiv:2109.04650, Sep 2021.
- C. Kim*, Heungsub Lee* et al., "torchgpipe: On-the-fly pipeline parallelism for training giant models," arXiv:2004.09910, Apr 2020.

*Contributed equally

Public Speeches

- "NSML, the hyper-scale ML training platform," KRnet, Jun 2022.
- "Remake of Hangulize," Golang Korea Meetup, Aug 2018.
- "Profiling," PyCon Korea, Aug 2015.
- "The server architecture of Durango," NDC, 2014, 2016, and 2018.

Languages

- Korean — Native
- English — Conversant in reading and writing

Education

Computer Software, Kwangwoon University, 2008, Completed the first year only.

(Last updated on Nov 22, 2023)