Seunghwan Hong

harrydrippin@gmail.com
github.com/harrydrippin / linkedin.com/in/harrydrippin
+82-10-4550-9287

Recent Professional Experiences

Scatter Lab (Pingpong Team) — ML Engineer (Seoul, Korea) Dec. 2019 - Present

Keywords: TensorFlow, PyTorch, Kubeflow Pipelines, Distributed Training (DeepSpeed), GCP, AWS, Faiss

- Implement and manage overall ML engineering parts including ML pipeline, serving optimization, data engineering, model optimization, internal tools/libraries.
- Build a pipeline for preprocessing and pseudonymizing 600+GB sized text data, and vector indexing using Kubeflow Pipelines.
    - Build a internal library that collects and manages filters for de-identifying data.
    - Build a pipeline for automatic build and deployment to manage Docker images for pipeline.
    - Build a research system on GCP that enables efficient research while maintaining privacy compliance.

- Optimize a pretraining process of large size language model for various models.
    - Optimize BERT pretraining process with distributed training strategies using 16-32 node cluster above multiple cloud components (Internal distributed training library, EFA, FSx, S3), collaborated with AWS MLSL.
    - Implement training code for training billion-size GPT-2 using DeepSpeed, and data preprocessing code using Apache Beam.
    - Conduct investigation for searching bottlenecks for optimizing Cloud TPU performance while pretraining using Cloud TPU Profiler.

- Conduct a research for multiple vector similarity search frameworks for real-time inference.
    - Build an early version of `faiss-serving`, server for inferencing vector similarity search above Faiss index using C++.
    - Refactor `faiss-serving` using multi-threaded worker on Python. Achieved 130 ~ 150 RPS with static memory usage above n-thousand concurrent users, which is 5x faster than early version.

- Implement initial version of Pingpong Flow (inference pipeline of 'Luda Lee', a conversational chatbot).
    - Build a library for loading MeCab on Java environment, enabling morpheme analysis with custom dictionary inside Spring Boot project. (github.com/scatterlab/mecab-ko-java)
    - Build a cloud-based log pipeline system to efficiently collect and statistically analyze the various types of logs from the chatbot pipeline and ML model using BigQuery and Cloud Logging.

- Build a Kubernetes cluster for deploying various internal tools, using Istio and Argo CD.
    - Build a model registry server using ML Metadata (TFX) and deploy to the internal cluster.

- Contribute to the establishment and settlement of an team development culture.
    - Build a team development guide for managing Python project, including contents about linter, CI/CD, commit convention, etc.
    - Lead various study sessions about Docker/Kubernetes and Go.

Common Computer (AI Network) — Software Engineer (Seoul, Korea) Sep. 2018 - Oct. 2019

Keywords: Docker, Typescript, Express.js, gRPC, Protocol Buffers, Firebase Realtime Database, AWS S3

- Built blockchain backend for executing a distributed computing job with ERC-20 token for enabling users to access our system more easily.
- Implemented gRPC, JSON-RPC based blockchain backend for making communication between Chrome Extension and distributed notes who are participating in the network as a computing node.

Hyprsense, Inc. — Software Engineer Internship (Burlingame, CA) Jun. 2018 - Aug. 2018

Keywords: Python, Typescript, Node, WebRTC, OpenCV, Redis, DynamoDB, Terraform

- Built framework for managing and annotating dataset including Valid Landmark, Tongue, and Face Checker, Misaligned Image Collector, Landmark Annotator, etc.
- Build deep learning model testing platform based on the web application by WebRTC connection between the web browser and inference server.
- Implemented interface linkage library for deep learning vision model and a web application by shipping lightweight model directly to the web browser. Video latency was reduced from 1~5 seconds to 6~10 milliseconds.
- Defined overall infrastructure as a code using Terraform and implemented Python tool for managing deployments within the CI/CD pipeline.

Technical Skills

Programming Languages: Python, Typescript (Javascript), C++, Go, Java, SQL
Frameworks / Platforms: TensorFlow, PyTorch, Docker, Kubernetes, Kubeflow, Flask, Express.js, FastAPI, gRPC, React.js, Terraform

Education

University of California, Irvine (Irvine, CA, USA) Jun. 2019 - Dec. 2019

Visiting Researcher in Informatics, Software Design and Collaboration Laboratory

Kookmin University (Seoul, South Korea) Mar. 2016 - Feb. 2020

Graduate in Computer Science, Average in Major: 3.5 / 4.0, 3.94 / 4.5

Recent Honors and Awards

Finalist, HACK/HLTH 2019 by AngelHack (Las Vegas, NV) Oct. 2019

Built a FHIR data pipeline for making medical data access permission system.

Finalist, F8 2019 Hackathon, Facebook (San Jose, CA) Apr. 2019

Build a chatbot and web application for tracking various problems on the city.

Prize for Best Engineering, 23rd Startup Weekend Apr. 2019

Build a chatbot and web application for tracking various problems on the city.

Recent Community Contributions and Invited Talks

AngelHack Seoul 2020 Online 2020

Organizer & Lead of Administration Team as a AngelHack Ambassador of Seoul, Korea.

Facebook Developer Circle: Seoul, F8 2019 Meetup 2019

Review Session for F8 2019 Hackathon: Invited Speaker