Joonbeom Park

Seoul, Republic of Korea.
joonb14@sktelecom.com
joonb14@gmail.com
jb.park@yonsei.ac.kr
github: https://github.com/joonb14

Experience

Infra & Backend & MLOps

Global Solutions Technology, SK Telecom, Pangyo, Korea
May 2025 ~ Present | Security AI Development Team

- Small Language Model(SLM) & Vision Language Model(VLM) serving with Triton | vLLM | Ray serve
- Data pre/post processing & prompting for the models with Airflow + FastAPI

shot up
January 2025 ~ April 2025 | Aug Team

- Designed Serverless infrastructure & implementation.
- Screenshot analyzing with AI Agent(LLM + VectorDB for RAG)
- Semantic search for saved screenshots.
- Deployed most of the AWS resources with AWS CDK(IaC).

aug: spacial social
September 2023 ~ December 2024 | Aug Team

- Designed Serverless event-driven infrastructure & implementation.
- Video streaming & image content delievery.
- GeoLocation-based AR Social Networking Service.
- Deployed most of the AWS resources with AWS CDK(IaC).

Global Solutions Technology, SK Telecom, Pangyo, Korea
February 2023 ~ December 2024 | Security AI Development Team

- Designed CVOps architecture(MLOps for AI surveillance cameras) & Implementation.
- Designed CVOps data pipeline & Implementation.
- Implemented image and data content delievery service.
- implemented model train, inference, conversion, deploy pipeline for surveillance cameras.
- Developed semantic image search.
- Co-Dev with AWS Engineers(AWS Fathom)
- Implemented batch inference API server for radio tower anomaly detection with drone images.
- Deployed most of the AWS resources with AWS CDK(IaC).

Software Engineer

Department of OS, Tmax Inc, Bundang, Korea
October 2022 ~ January 2023 | Online Meeting Team

- Implemented C++ libtorch denoiser based on python PyTorch denoiser for Online Meeting solution based on Open WebRTC Toolkit(OWT) media server.

June 2022 ~ September 2022 | Virtual Desktop Solution Team

- Documentation of Tmax Virtual Desktop Solution(VDS)'s common library.
- Developed RTP/SRTP packet sender for WebRTC testing.

Game Server Developer

Department of Metaverse, Tmax Inc, Bundang, Korea
December 2021 ~ May 2022 | Metaverse Server Team

- Developed game server using Libuv(Node.js core c++ library).
- Developed Unity client to test Libuv game server.

Machine Learning Engineer

Department of AI, Tmax Inc, Bundang, Korea
September 2021 ~ November 2021 | Computer Vision Team

- Developed realtime pose estimation with Unity 3D avatar on smartphones using BlazePose(Google Mediapipe).

May 2021 ~ June 2021 | Computer Vision Team

- NVIDIA Triton server inference system maintenance

February 2021 ~ August 2021 | Computer Vision Team

- Granted KISA(K-NBTC) BIO Authentication test(ISO/IEC JTC1 SC37 standard) certification with face recognition algorithm HyperFace.
- Developed realtime face recognition Android application using Google Android ML Kit.

Graduate Researcher

Department of Computer Science, Yonsei University, Seoul, Korea
March 2019 ~ February 2021 | Mobile Embedded System Lab | Advisor: Professor Hojung Cha

- Developed a gaze data collecting application for Android using Google Android ML Kit.
- Developed a convolution neural network for realtime gaze estimation model on smartphones using Tensorflow Lite.
- Developed an Android application for realtime gaze estimation with front facing camera on smartphones using Google Android ML Kit.

Undergraduate Researcher

Department of Computer Science, Yonsei University, Seoul, Korea
December 2017 ~ February 2019 | Mobile Embedded System Lab | Advisor: Professor Hojung Cha

- Developed a SVM model for estimating power consumption of smartphone(Pixel XL) display based on the RGB values of screen display.
- Developed energy aware UI design tool utilizing the SVM model, and image clustering.
- Developed iBeacon logger application for Android, and visualization tool on web browser using Highcharts and GoJS.

Paper

GAZEL: Runtime Gaze Tracking for Smartphones

2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)
DOI: 10.1109/PERCOM50583.2021.9439113
J. Park, S. Park, H. Cha, "GAZEL: Runtime Gaze Tracking for Smartphones," *The 19th International Conference on Pervasive Computing and Communications (PerCom 2021), Virtual Conference*, Mar. 22-26, 2021.

Projects

Synapsego: Automated Visual Synopsis Creation, SK Telecom
May 2025 ~ Present

- Audio transcription model serving with Ray Serve
- VLM multimodal inference serving and SLM serving with vllm + FastAPI
- Pipelining with Airflow DAGs

Shot Up: AI Assistant for Screenshots, with Gunhee Han, Harry Kim, Hojin Joo
January 2025 ~ April 2025
AppStore Link

- Design and deploy the AWS resources with AWS CDK.
- Firebase authentication with Google & Apple login.
- Mobile push notification with AWS SNS, Firebase Cloud Message(FCM).
- Business logic with Lambda Function URL(to support streaming responses) + CloudFront distribution
- Lambda@Edge authentication for CloudFront origins
- OpenSearch Serverless vector collection and AWS Bedrock & ChatGPT for semantic search

Aug: Location-based AR Social Network Service, with Gunhee Han, Harry Kim, Hojin Joo
September 2023 ~ December 2024
AppStore Link

- Design and deploy the AWS resources with AWS CDK.
- Firebase authentication with email & password, Google login, Apple login.
- Mobile push notification with AWS SNS, Firebase Cloud Message(FCM).
- Video streaming & image content distribution with AWS Route53, CloudFront, S3.
- Automatic thumbnail creation with AWS S3, Lambda trigger.
- Automatic HLS convertion for videos with AWS S3, Lambda trigger, MediaConvert.
- Business logic with Event-driven architecture. Using only serverless resources, AWS REST API Gateway, EventBridge, Lambda, DynamoDB.

- Geolocation-based content search API based on Mapbox quadkey.

MLOps pipeline for AI Surveillance Cameras(CVOps), SK Telecom
February 2023 ~ September 2024

- Design and deploy the AWS resources with AWS CDK.
- User Authentication with AWS Cognito.
- Dataset management with AWS Fathom(co-developing service with SKT) SDK.
- Image preview and data streaming middleware service with AWS CloudFront + ALB + Fargate + Fathom SDK
- Training pipeline with Sagemaker TrainingJob. Progress update, logging implemented with AWS SQS, Lambda.
- Inference pipeline with Sagemaker ProcessingJob. Progress update, logging implemented with AWS SQS, Lambda.
- Implementing Conversion pipeline, Torch to ONNX to binary file conversion. Each feature uses SageMaker ProcessingJob for conversion. Orchestrating workflow with Step Function.
- Pipeline(Train, Inference, Conversion) implementation with AWS Step Function.
- Model deployment API with AWS API Gateway, ALB, Fargate, FastAPI. Updated AI model download logic integration with S3 presigned URL.
- Semantic image search with AWS OpenSearch. Feature extraction with OpenCLIP on Fargate.

Anomaly detection(with drone images) service for radio towers, SK Telecom
June 2023 ~ August 2023

- Whole AWS infrastructure with AWS CDK(IaC).
- VPC & NAT Gateway for security requirements.
- Batch inference anomaly detection model with AWS Sagemaker ProcessingJob to only use instance on-demand(Serverless Service).
- Handling non-code level errors such as AWS Sagemaker resource errors with AWS Step Functions. Triggering AWS Lambda if errors occur, then report it to the server.

Porting python denoiser module to Tmax Online Meeting solution, Tmax OS
October 2022 ~ December 2022

- Converting PyTorch denoiser model to C++ torchscript model which is customized from FaceBook denoiser project.
- Converting python denoiser inference script to C++ libtorch code.
- Applying the model to Node.js Tmax Online Meeting based on Open WebRTC Toolkit Media Server (On progress).

Documentation and testing for Virtual Destktop Soultion (VDS), Tmax OS
June 2022 ~ September 2022

- Writing guidelines for using VDS common library, especially for establishing TCP/WebSocket Channel connection, and Signaling interface in VDS.
- For testing VDS server, developed RTP/SRTP media packet sender.

Game server development using Libuv, Tmax Metaverse
December 2021 ~ May 2022

- Developing C++ Metaverse Game Server with Libuv(Node.js core library) for C# Unity Client. Demo.
- **TCP** Server for realtime multiplayer games.
- Designing packet protocol for Metaverse server and Unity cleint.
- Handling 100+ players in a **stateful** Metaverse server.

Realtime 3D pose estimation on smartphones with Unity, Tmax AI
September 2021 ~ November 2021

- Developed realtime pose estimation with Unity 3D avatar on smartphones using BlazePose(Google Mediapipe) with Unity Barracuda.
- Modified BlazePoseBarracuda & ThreeDPoseUnityBarracuda for development.
- Optimized application to meet 16 FPS on Galaxy S21.
- Demo.

NVIDIA Triton Server Maintenance, Tmax AI
May 2021 ~ June 2021

- Deployed our team's deep learning models with NVIDIA Triton Server on Tmax HyperCloud(customised k8s)
- (Deprecated) Used Flask gateway for pre & post processing
- Modified input stream format. JSON to byte buffer
- Changed pre & post processing to use Ensemble + Python BLS pipeline

Face recognition algorithm development, Tmax AI
February 2021 ~ August 2021

- Implemented face recognition algorithms on TensorFlow 2
- Converted TensorFlow 2 model to TensorFlow Lite
- Applied uint8 quantization on TensorFlow Lite model.
- Developed realtime face recognition PC/Android application using OpenCV/Android ML Kit and IJB-C dataset. Android Demo.
- Further developed *HyperFace* face recognition algorithm to get KISA(K-NBTC) BIO Authentication test(ISO/IEC JTC1 SC37 standard) certified.
- Developed *Face Bird* game application with *HyperFace*. Demo.

Realtime gaze estimation on smartphones, Yonsei University
March 2019 ~ February 2021 | Advisor: Professor Hojung Cha

- Developed realtime eye region Bitmap cropper, and landmark collector with Google Android ML Kit.
- Developed button click based auto gaze data collecting application for Android using Google Android ML Kit.
- Developed Android laucher application for gaze data collection.
- Developed a light-weight convolution neural network for realtime gaze estimation model on smartphones using TensorFlow Lite.
- Used tablet gaze data for training, applied linear regression to use this model on smartphones.
- Utilized RenderScript for converting RGB image to 1 channel Black & White image.
- Developed an Android application for realtime gaze estimation with front facing camera on smartphones using Google Android ML Kit.
- *GAZEL: Runtime Gaze Tracking for Smartphones* paper publication.

- Developed *Gaze Bird* game application with *GAZEL*. Demo.

Power management on embedded systems, Yonsei University
December 2017 ~ February 2019 | Advisor: Professor Hojung Cha

- Developed a SVM model for estimating power consumption of smartphone(Pixel XL) display based on the RGB values of screen display using Monsoon ADB.
- To decrease power consumption on smartphones, developed *Energy Aware UI Design Tool* which runs on Flask. ver1. ver2.
- Developed iBeacon logger application for Android, and visualization tool on web browser using Highcharts and GoJS.

Social

Teacher at Daesung high school, Daesung high school Seoul, Korea.
September 2017 ~ December 2017, Teacher

- Taught C++ programming for high school students on weekends.

Education

M.S in Computer Science, Yonsei University (February 2021)
B.S in Computer Science, Yonsei University (February 2019)

Skills

Dev. Languages: AWS, CDK, Python, TypeScript, C++, Java(Android Studio)