Jun-Min Lee

KAIST, Seoul, South Korea | Mail: ljm56897@gmail.com | Website: ljm565.github.io

LinkedIn: jun-min-lee-189383264 | GitHub: github.com/ljm565 | Portfolio (On-line)

Research Interests

I am an AI developer passionate about using artificial intelligence to make the world a better place. Inspired by the rapid progress of technologies like LLMs and MLLMs, I'm excited by AI's growing ability to perform tasks once limited to humans. This has deepened my interest in LLMs, which I now study and work with. I strive to gain diverse experiences to become an AI researcher and engineer who creates meaningful impact through AI.

- Healthcare NLP
- Generative model (e.g., LLM, multi-modal LLM)
- LLM (including PEFT) and deployment

Skills

- Programming: Python, Java (Spring), JavaScript, FastAPI
- Framework & Tools: PyTorch, PyTorch Lightning, LLM Training & deployment (Triton-client, TensorRT-LLM, TensorRT-LLM Backend, vLLM, ollama), PEFT, RAG, Docker, Git, Elasticsearch, GCP
- Environment preferences: Linux, Mac
- Others: PostgreSQL, HTML, CSS

Education

Ph.D. student in Graduate school of AI, KAIST (Advisor: Prof. Edward Choi) Sep. 2021 - present

- I fulfilled my three-year military service obligation by working at a software development company.

M.S. in Aerospace Engineering, KAIST Sep. 2019 - Sep. 2021

B.S in, Aerospace Engineering (Double major: Industrial Design), KAIST Mar. 2015 - Sep. 2019

Work Experience

Machine Learning Engineer (Student intern), Kakao Healthcare (DS/IX team) Oct. 2024 - Apr. 2025

- Performed both PEFT and full training of a medical NER-specialized LLM, and deployed the model to multiple specialized medical centers using vLLM for inference.
- Built a GCP-based healthcare counselor chatbot by integrating RAG to provide accurate and context-aware health consultations.

Machine Learning Engineer (Military service replacement), Lomin (ML team) Apr. 2023 - Sep. 2024

- Trained a document-specialized LLM for document understanding tasks and deployed it on-premises to enterprise environments using TensorRT-LLM and TensorRT-LLM Backend for optimized inference.
- Deployed various deep learning models—including text detection, recognition, and classification—to multiple enterprises in on-premises environments using BentoML and Triton client.
- Customized YOLOv8 to improve the performance, speed, and robustness of the text detection model specifically for document images.
- Built a zero-shot document classification system by developing a document embedding model capable of generalizing to unseen document types without task-specific fine-tuning.
- Departed from the conventional two-stage OCR approach by designing and implementing an end-to-end pipeline that unifies text detection and recognition, resulting in enhanced management efficiency.
- Developed a lightweight Autoscan model capable of real-time prescription detection, and deployed it using TensorFlow.js for in-browser inference across multiple pharmacies without the need for server-side computation.

Machine Learning Engineer (Military service replacement), IBRICKS (AI Tech) Sep. 2021 - Apr. 2023

- Maintained and added new features to a document analysis product developed in Java, ensuring its stability and

continuous improvement in functionality.

- Designed and implemented an open-domain chatbot system using models like GPT-2, DialoGPT, and BART prior to the rise of LLMs.
- Built a BERT-based extractive summarization model for long documents by identifying and extracting key sentences, enabling concise and relevant summaries.
- Reduced model size and complexity by replacing traditional sequence-to-sequence architecture with a shared encoder-decoder model, improving efficiency while maintaining performance.
- Developed TESGAN (Text Embedding Space GAN), an unsupervised GAN-based model that generates rich embedding space for text data augmentation, designed prior to the advent of LLMs.
- Developed an OCR model specialized for extracting text from online home shopping advertisement images.

Research Contributions

- 2025 PC Member, Association for the Advancement of Artificial Intelligence (AAAI)
- 2023, 2024 PC Member, Association for the Advancement of Artificial Intelligence (AAAI)
- 2023 Industry Track Committee, Empirical Methods in Natural Language Processing (EMNLP)
- 2023 PC Member, Empirical Methods in Natural Language Processing (EMNLP)
- 2023 PC Member, Association for Computational Linguistics (ACL)

Publications

- **Jun-Min. Lee** and Tae-Bin Ha (2023), Unsupervised Text Embedding Space Generation Using Generative Adversarial Networks for Text Synthesis, *Northern European Journal of Language Technology (NEJLT), 9(1)*. [PDF]
- **Jun-Min Lee***, Hyun-Soo Kim*, Tae-Bin Ha, Hojin Park, and Youngmin Ahn (2021), Open-Domain Dialogue Generation using Pre-trained Language Models in Korean, *Conference of Korea Computer Congress (KCC)* (*: equal contribution). [PDF]
- **Jun-Min Lee**, Yunshil Choi, and Jung-Ryul Lee (2022), Laser structural training, artificial intelligence-based acoustic emission localization and structural noise signal distinguishment in a thick FCEV fuel tank, *International Journal of Hydrogen Energy (IJHE), 47(6), 4236-4254*. [PDF]
- **Jun-Min Lee**, and Jung-Ryul Lee (2021). Acoustic emission localization on composite hydrogen storage tank and feature analysis of acoustic emission and noise signals. *Conference of The Korean Society for Aeronautical and Space Sciences (KSAS)*. [PDF]
- Yunshil Choi, **Jun-Min Lee** and Jung-Ryul Lee (2020), Nondestructive Testing and Structural Health Monitoring for Pressure Vessels of FCEV using Guided-Wave Ultrasonic Propagation Imager, *Conference of The Korean Society for Composite Materials (KSCM)*.
- **Jun-Min Lee**, Yunshil Choi, and Jung-Ryul Lee (2020), Structural Health Monitoring of Hydrogen Pressure Vessel using Artificial Intelligence, *Conference of The Korean Society for Nondestructive Testing (KSNT)* (**Award**).
- **Jun-Min Lee**, Yunshil Choi, and Jung-Ryul Lee (2019), Acoustic Emission Simulation using Pulse Laser-induced Wave with Considering Arrival Time for Quantification, *The 1st Korea-China-Japan Joint Symposium on Composite Materials*.

Projects

Universal LLM Trainer (Link: github.com/ljm565/universal-llm-trainer)

- A modular training framework for easily fine-tuning open-source LLMs. Supports instruction tuning, full fine-tuning, LoRA, QLoRA, and training strategies like DDP, FSDP, and gradient checkpointing. Enables autoregressive pre-training and SFT-based instruction tuning with simple model wrapper integration.
- Tools Used: Python, PyTorch