Yunjae Won

yunjae.won@kaist.ac.kr | yunjae-won.github.io

Education

KAIST Graduate School of AI Seoul, Republic of Korea
Integrated M.S./Ph.D., Advisor: Minjoon Seo *Fall 2024 – Present*

KAIST Daejeon, Republic of Korea
Bachelor of Science in Electrical Engineering, Minor in Computer Science *Feb. 2018 – Aug. 2024*
GPA: 3.76/4.3

Publication

Differential Information: An Information-Theoretic Perspective on Preference Optimization
Yunjae Won, Hyunji Lee, Hyeonbin Hwang, Minjoon Seo *May 2025*
Preprint. Under Review. *arXiv*

Experience

Machine Learning Engineer Intern Seongnam, Republic of Korea
NAVER G Place AI Development *Mar. 2023 – Aug. 2023*

- Developed an easy-to-use distributed hyper-parameter optimization tool for a Kubernetes-based environment.
- Trained a light-weight accurate Image Quality Assessment model, reducing the model's size by over 10x while improving accuracy by 56%; model has been in-service since July 2023 and enabled a shift from GPU to CPU-based serving.

Machine Learning Engineer Intern Seongnam, Republic of Korea
NCSOFT Speech AI Lab *July 2022 – Aug. 2022*

- Developed an in-the-wild audio signal preprocessing pipeline for training Singing Source Extraction Models.
- Increased training data set size twelve-fold and trained a new model that outperformed the previous state-of-the-art, improving the Signal-to-Distortion Ratio from 8.06 to 8.45.
- Awarded 'Excellent NCSOFT Summer Internship Project' and offered a 6-month extension.

Undergraduate Research Assistant Daejeon, Republic of Korea
Data Intelligence Lab, KAIST School of Electrical Engineering *Jan. 2023 – Feb. 2023*

- Investigated new methods for incorporating contrastive learning techniques for active learning.
- Measured the performance degradation of popular Active Learning methods under instance-dependent and class-dependent label noise on the MNIST dataset.

Military Service Seongnam, Republic of Korea
Republic of Korea Air Force *July 2020 – Feb. 2022*

- Developed a Visual Basic Script for battalion level personnel management.

Projects

A Closed-Form Expression for Unalignment *Spring 2025*

- Project for KAIST AI707 <Advanced Topics in Deep Reinforcement Learning>
- Derived a closed-form expression for the ideal distribution of rejected responses under the DPO framework, showing the possibility of generating harmful responses from aligned policies.

Orthogonal Gradient Descent: Learning from Preferences with Minimal Forgetting *Fall 2024*

- Project for KAIST AI611 <Deep Reinforcement Learning>
- Proposed a method using projected gradient descent to learn from human preferences while minimizing the forgetting of previous knowledge.

Metabolic Reaction Prediction via Next Token Prediction *Fall 2024*

- Project for KAIST AI607 <Graph Mining and Social Network Analysis>
- Formulated the metabolic reaction prediction task as a next token prediction problem, proposing a transformer architecture to predict the next reaction.

Korean Text Recognition Challenge | *1st Place out of 1,158 participants* *Jan. 2023*

- Developed a Korean text recognition model on a dataset of handwritings from Korean children for the Kyowon Group AI Challenge.
- Outperformed an OCR corporate team backed by state-of-the-art GPUs using only limited Google Colaboratory resources.

Agricultural Products' Price Change Forecasting Challenge | *3rd Place out of 705 participants* *Sep. 2022*

- Constructed an Extra-Trees Regressor based time-series forecaster for a competition hosted by the Korea Agro-Fisheries & Food Trade Corporation.
- Devised a new data-augmentation technique using a polynomial trend based pseudo-labeling process.

Korean Face Open Set Verification *Fall 2022*

- Project for KAIST EE488 <Deep Learning for Computer Vision>
- Trained an EfficientNet-based embedding network using metric learning on a crowd-sourced Korean celebrity face image dataset.
- Achieved the best model performance among models trained without any additional training data, leading to an invitation from the professor to present the development process.

Research Interests

- Preference Optimization, Reinforcement Learning from Human Feedback
- Continual Learning, Knowledge Distillation
- Optimization
- Large Language Models