Ahn Young Jin (He/Him)

github.com/snoop2head

EDUCATION

Korea Advanced Institute of Science and Technology (KAIST)   Seoul, South Korea
M.S. in Artificial Intelligence   *September 2023 –*
• Advisor: Kee-Eung Kim

Yonsei University   Seoul, South Korea
B.A. in Economics & Minor in Applied Statistics   *March 2015 – August 2023*
• GPA: 3.81 / 4.3 overall, [ECON] 3.78/4.3, [STATS] 3.89/4.3
• Relevant courseworks: [STATS] Mathematical Statistics A+, R and Python Programming A+
 [CS] Theory and Practice of Deep Learning A+, Data Structures B+

TECHNICAL SKILLS

Languages Python, R, JavaScript, Rust (* in the order of fluency)
Libraries PyTorch, Pytorch Lightning, Huggingface, Jax, Flax, Haiku, OpenCV, Docker, Django

LANGUAGE SKILLS

Korean Native.
English Fluent, TOEFL iBT 110/120, Served as KATUSA platoon leader in USANEC Casey.
Chinese Novice, HSK 4.

RESEARCH EXPERIENCES

Principal Investigator   September – November 2022
Ministry of Science and ICT funded Private Data Neural Filtering National R&D Project
• Managed ₩200,000,000 research fund and research participants.
• Provided data labelling companies with annotation guidelines of privacy benchmark dataset.

Research Intern   July – August 2022
Cyber Security Lab. Department of Computer Science, College of Computing, Yonsei University
• Implemented membership inference attack methods on Large Language Models and CNNs.

Research Intern   January – April 2023
Vision and Learning Lab. Department of AI, College of Computing, Yonsei University
• Devised joint classification approach of time-aligned quantized tokens on visual speech recognition.
• Achieved SOTA performance in Lip Reading in the Wild benchmark.

PROJECT EXPERIENCES

KoDALLE: Text to Fashion | Github Repo   November 2021 – January 2022
Generating dress outfit images based on given input text
• Resolved scarcity of target language dataset with pretrained language model's embedding.
• Preprocessed and maintained versions of 1 million pairs of image-caption dataset.
• Trained model from scratch for VQGAN and DALLE.

DotNeuralNet & DotT5 | [Github Repo](#) | [Demo](#)   March 2023
End-to-end optical braille recognition to braille pattern translation
• Generalized over non-existent braille signs by one hot encoded multi label classification.
• Trained light-weight object detection model for braille recognition in natural scenes.
• Built braille to text conversion T5 model with BLEU score of 75%.

ElimNet | [Github Repo](#)   October 2021
Eliminating top layers in a neural network pretrained with large dataset for downstream task.
• Replaced time-consuming and underperforming NAS models with pruned CNN models.
• Removed top layers from pretrained CNN to create lightweight model less than 1M #params.
• Saves inference time by 20% while maintaining or improving its performance.

Gomoku-GPT | [Demo](#)   April 2023
GPT trained with curriculum learning consisted of MCTS and supervised learning for Gomoku game.
• Trained GPT from scratch as policy neural network in order to eradicate CNN's problem of sparse feature extraction and lack of sequential data processing.
• Complemented reinforcement learning's asymmetry problem with supervised learning.

KoGPT-Joong-2 | [Github Repo](#) | [Demo](#)   December 2021
Emotional prompts for 1.2B #params GPT model finetuned with lyrics and poetry dataset.
• Generated sentences and acrostic poems based on emotional prompt.
• Maintained huggingface hub demo hosting GPT model with LM head.
• Finetuned KoGPT-Trinity with huggingface trainer.

Realtime Turtleneck Alert | [Github Repo](#) | [Demo](#)   November – December 2020
Real-time desk posture detection through webcam.
• Converted image dataset into Yaw/Pitch/Roll numerical dataset using RetinaFace model.
• Trained and hypertuned random forest classification model with precision rate of 93%, outperforming CNN models by +20%p margin.

COMPETITIONS

NIPA Machine Reading Comprehension Competition | [Github Repo](#)   June - July 2022
(2nd place / 26 companies) Question answering task of extracting answers from given context.
• Specified metrics into start token prediction accuracy and end token prediction accuracy.
• Customized model head with multi sample dropout, LSTM head and CNN head.
• Replicated performance in low computing resource with FP16 and gradient accumulation.

Ministry of Statistics Korean Standard Industry Classification   April 2022
(7th place / 311 teams) Sequence classification task for classifying 200 categories and sub-categories.
• Compared across publicly available pretrained language models' performances.
• Constructed train and inference pipeline in pytorch lightning.

Dacon KLUE Natural Language Inference Competition | [Github Repo](#)   January 2022
(1st place / 468 teams) Utilizing R-BERT and Improved Baseline architecture from TACRED.
• Applied typed-marker method introduced in Zhou et al(ACL-IJCNLP 2022) and selective sentence pooling approach from Wu et al(CIKM 2019) for the given NLI task.
• Composed custom dataset, trainer, and inference pipeline in pytorch.

Dacon Code Clone Detection Competition | [Github Repo](#)   June 2022

(3rd place / 337 teams) Identifying code similarity with encoders from seq2seq model.
• Solved token length constraint by dropping code blocks, removing comment lines and filtering out unused function within the code.
• Implemented regularization method for input sequence order with KL-Divergence loss referencing Wu et al(NeurIPS 2021).

Dacon Stock Price Prediction Competition | [Github Repo](#)   September 2021

(6th place / 205 teams) Predicting the Closing price of 376 Korean public companies.
• Validated the model's performance on ARIMA, RandomforestRegressor and ElasticNetCV.
• Derived financial data utilizing TA-Lib and FinanceDataReader.

Geulstagram Dataset | [Github Repo](#)   September 2019 – March 2020

3rd place for Selectstar & National Information Society Agency(NIA) dataset proposal competition.
• Managed pull request / version control for emotional Instagram posts(Geulstagram) dataset.
• Converted Korean image file to text file using Google Cloud Vision API.

OPENSOURCE CONTRIBUTION

docker/docker.github.io | [Pull Request](#)   March – April 2020

Source repo for Docker's Official Documentation.
• Updated PostgreSQL initialization for "Quickstart: Dockerizing Django" documentation.

NVlabs/stylegan2-ada-pytorch | [Pull Request](#)   November 2021

Source repo of NVIDIA Research for StyleGAN2-ADA.
• Fixed torch version comparison fallback error.

WEB PROJECTS

yonsei.exchange | [Github Repo](#) | [Website](#)   July 2021 –

Studying abroad summaries for Yonsei students: 22000 Pageviews & Average Retention Time 3 min+
• Collected and preprocessed 11200 text review data from the Yonsei website.
• Visualized department distribution and weather information using matplotlib.
• Sentiment analysis on satisfaction level for foreign universities with pretrained BERT model.
• Clustered universities with provided curriculum with K-means clustering.

FitCuration | [Github Repo](#)   February – April 2020

Search-based exercise retrieval web service.
• Built retrieval algorithm based on search keyword using TF-IDF.
• Deployed website using Docker, AWS RDS, AWS S3, AWS EBS.
• Constructed backend using Django, Django ORM & PostgreSQL.

HONORS AND AWARDS

Academic Honors, Yonsei University   *Spring, 2015*

LEADERSHIP AND SERVICE

Academic support for North Korean refugee students, 80+ hours   *September 2015 – April 2016*
Yonsei Vision Up academic support for high school students   *March – December 2016*
Military Service for Republic of Korea, Senior KATUSA   *September 2017 – May 2019*