KIYOON YOO

+(82)10-2293-0499
961230@snu.ac.kr ⋄ Google Scholar ⋄ Github

RESEARCH INTEREST

- Adapting LM to other domains
  - Conversational AI [I1]
  - Medical domain [W2, ArXiv1]
- Safety in Language Models
  - Watermarking texts and language model outputs [C1, C3]
  - Adversarial attack and defense, backdoor attacks [C5, C6]
- Miscellaneous
  - Model compression [C2,C7,ArXiv3]
  - Representation learning [C4, W3]

INTERNSHIPS

- **Naver Cloud, AI Lab (Research Intern)** May 2024 -
  - Advisor: SangDoo Yoon
- **Naver Webtoon (Research Intern)** Aug. 2023 - Dec. 2023

  I1 Conversational AI
  - Data pipeline: Data generation - curation - finetuning via API
  - Evaluation of models in hallucinations
  - Preprocessing of public korean corpus and pretraining sLLM

  I2 Research in watermarking LLMs and copyright protection for T2I generative models
  - Led research on watermark for LLMs and LLM output detection (1 paper accepted to NAACL 24)
  - Collaborated on a project for protecting against style imitation
  - 3rd place in "N Innovation 2023" (Competition for innovative business & services across Naver)
- **Naver Webtoon (Research Intern)** June 2022 - Aug. 2022

  I3 Research in natural language watermarking (1 paper accepted to ACL 23)

EDUCATION

Ph.D. Candidate in Intelligence and Information,
Department of Intelligence and Information, Seoul National University Sept. 2019 - Present
Supervised by Prof. Nojun Kwak

BS in Food Science and Biotechnology,
Department of Food and Animal Biotechnology, Seoul National University March 2015 - Aug. 2019
Minor in Industrial Engineering

PUBLICATIONS

<International Conference>

C1 **KiYoon Yoo**, Wonhyuk Ahn, and Nojun Kwak. “Advancing beyond identification: Multi-bit watermark for large language models.”, NAACL 2024. (long paper, oral presentation)

C2 Jangho Kim, Jayeon Yoo, Yeji Song, **KiYoon Yoo**, Nojun Kwak,“Dynamic Collective Intelligence Learning: Finding Efficient Sparse Model via Refined Gradients for Pruned Weights”, The 31st ACM International Conference on Multimedia (ACM MM 2023).

C3 **KiYoon Yoo**, WonHyuk Ahn, Jiho Jang, Nojun Kwak. “Who Leaked My Document? Robust Natural Language Watermarking through Invariant Features”, ACL 2023 (long paper).

C4 Jiho Jang, **KiYoon Yoo**, Seonhoon Kim, Kong Chaerin, Jangho Kim, Nojun Kwak, “Self-Distilled Self-Supervised Representation Learning”, WACV 2023.

C5 **KiYoon Yoo**, Nojun Kwak, “Backdoor Attacks in Federated Learning by Rare Embeddings and Gradient Ensembling”, EMNLP 2022 (long paper, oral presentation)

C6 **KiYoon Yoo**, Jangho Kim, Jiho Jang, Nojun Kwak, “Detection of Word Adversarial Examples in Text Classification: Benchmark and Baseline via Robust Density Estimation”, Findings of ACL 2022 (long paper).

C7 Jangho Kim, **KiYoon Yoo**, Nojun Kwak, “Position-based Scaled Gradient for Model Quantization and Sparse Training”, Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS 2020).

<International Journals>

J1 Hojun Lee, Donghwan Yun, Jayeon Yoo, **KiYoon Yoo**, Yong Chul Kim, Dong Ki Kim, Kook-Hwan Oh, Kwon Wook Joo, Yon Su Kim, Nojun Kwak, Seung Seok Han, ”Deep Learning Model for Real-Time Prediction of Intradialytic Hypotension”, Clinical Journal of the American Society of Nephrology, 2021.

<Workshops>

W1 **KiYoon Yoo**, Wonhyuk Ahn, Yeji Song, and Nojun Kwak. “Exploring Causal Mechanisms for Machine Text Detection Methods.”, Workshop on Trustworthy Natural Language Processing@NAACL 2024

W2 Hyeonjin Kim, Min Kyu Kim, Jae Won Jang, **KiYoon Yoo**, Nojun Kwak, ‘TEAM MIPAL at MEDIQA-M3G 2024: Large VQA Models for Dermatological Diagnosis”, *2nd Place (English) for Multilingual & Multimodal Medical Answer Generation*, Workshop on ClinicalNLP@NAACL 2024

W3 Inseop Chung, **KiYoon Yoo**, Nojun Kwak. ”Open Domain Generalization with a Single Network by Regularization Exploiting Pre-trained Features.”, Workshop on Data-centric Machine Learning Research@ ICLR 2024.

W4 **KiYoon Yoo**, Nojun Kwak, “Backdoor Attacks in Federated Learning by Rare Embeddings and Gradient Ensembling”, Workshop on Federated Learning for Natural Language Processing@ACL 2022.

<Work in Progress>

ArXiv1 **KiYoon Yoo**, Min Kyu Kim, Jae Won Jang, Hyeonjin Kim, Nojun Kwak, “Synthetic Generation of Multimodal Patient-Doctor Conversation for Skin Lesion Diagnosis”, Non-archived.

ArXiv2 NamHyuk Ahn, WonHyuk Ahn, **KiYoon Yoo**, DaeSik Kim, SeungHoon Nam. “Imperceptible Protection against Style Imitation from Diffusion Models.”, ArXiv preprint.

ArXiv3 Geunjae Choi, Kamin Lee, **KiYoon Yoo**, and Nojun Kwak. ”Hardware-Friendly Post-Training Quantization: Input-and Output-Channelwise Scale and Offset.” ArXiv preprint.

PROJECTS

Building Korean Text Classification and Question Answering Models March 2022 - Nov. 2022

- Funded by WIGO
- Training NLP models for text classification and question answering tasks (KorQuad 2.0)

Partial Discharge Detection by Artificial Intelligence-based Algorithm Sept. 2020 - June. 2021

- Funded by LS Cables and Systems
- Build neural network for anomaly detection on underground cables
- Preprocessed and trained on real world data

Real-time / Lightweight Object Detection in Embedded Systems Jan. 2020 - Dec. 2021

- Funded by Samsung Electronics
- Research project on efficient training (quantization and model pruning on classification models)

ACADEMIC SERVICE

Reviewer

- ARR, EMNLP,ACL, ECCV, CVPR
- Neurocomputing

Program Committee

- Trustworthy Natural Language Processing@ACL2023