KUEI-CHUN KAO

• *linkedin.com/in/kuei-chun-kao* • *Github* • *Personal Website* • *(310) 621-1492* • *johnson0213@g.ucla.edu*

EDUCATION

University of California, Los Angeles (UCLA) *Los Angeles, CA*
M.S. in Computer Science, Specialization in Artificial Intelligence, Overall GPA: 3.9/4.0 *Sept. 2023 – Present*
Graduate Student Researcher including scholarship advised by Prof. Cho-Jui Hsieh

National Yang-Ming -- Chiao-Tung University (NYCU) *Hsinchu, Taiwan*
B.S. in Computer Science, Overall GPA: 3.84/4.0 *Sept. 2018 – Dec. 2022*
Research Assistant advised by Prof. Yung-Ju Chang with Stanford Screenomics Lab, Prof. Wei-Chen Chiu, and Prof. Keh-Yih Su
Teaching Experience: *Principal Teaching Assistant of Introduce to Natural Language Preprocessing (2022 Spring)*

WORK EXPERIENCE

Himax Imaging *Irvine, CA*
AI Engineer Intern *June. 2024 – Sept 2024*

- Developed and trained a unified model via Yolov8n backbone for face detection, emotion recognition, and head pose estimation, achieving a **30%** reduction in inference time and model size through model compression techniques such as knowledge distillation and quantization.
- Deployed the optimized model onto edge device AI chips, enabling real-time metadata processing on Window's notebooks and integrated seamless communication with Microsoft Copilot, enhancing the efficiency and performance of AI-driven applications within 3 months.

Appier *Taipei, Taiwan*
Machine Learning Scientist (Real Team Bidding Team on Digital Advertisement) *Sept. 2021 – July 2023*

- Improved User Lookalike models to produce the precise user score for each unique user ID based on existing client site activities and deployed models in production using CI/CD pipelines, enhancing **20%** improvement on AUROC compared to baseline.
- Extracted user behavior patterns from **1000K+** conversion funnel data and analyzed the Click-Through Rate of different campaigns and supply side platforms using PySpark, SQL, and Pandas, resulting in **120%** revenue and Cost-per-Click growth within 3 months.
- Improved robustness of bidding models through calibration and auto-tuning techniques, enhancing stability on unseen impression data.
- Developed Grafana dashboards for monitoring and alerting of outlier training data, saving up **30%** of the model troubleshooting time.
- Integrated and refactored different data serving pipelines via Spark and Kubernetes, causing **40%** of data pipeline storage reduction.

Cinnamon AI *Taipei, Taiwan*
AI Engineer Summer Intern *July. 2021 – Sept. 2021*

- Implemented and trained a Seq2Seq-based trip advisor model on Gradio.io, designed to recommend tourist attractions and optimize trip routes based on customer preferences, streamlining the process of model serving and enhancing user interaction through intuitive interface.

SELECTED PROJECTS AND PUBLICATIONS

Auto-Prompting Design for LLMs and MLLMs ***[EMNLP'24] [code] [website]*** *Sept. 2023 – Present*

- Proposed an auto-prompting approach to synthesize large-scale multi-image geometry problems based on different skills and utilized supervised fine-tuning to enhance open-source MLLMs reasoning ability on widely used multi-image benchmark. *(Under Submission)*
- Proposed an auto-prompting approach that combines LLMs and external symbolic solver to solve Algebra Word Problem by utilizing different prompting strategies, achieving **30%** improvement in answer accuracy on both English and Chinese datasets.
- Curated a new and larger algebraic dataset with prompt optimization which contains multiple variables questions to evaluate our proposed reasoning approaches on solving more challengeable Algebra Word Problem.

Predicting Smartphone Users' Kill Time Moments ***[Ubicomp'21, CHI'23, preprint (CHI'25)] [code1, code2]*** *July 2020 – Present*

- Investigated the use of MLLMs, in the analysis of smartphone user activities from screenshots, offering an alternative to traditional app usage data analysis limitations, providing highly convinced reliability score compared to human coders. *(Under review CHI'25).*
- Leveraged deep learning fusion model to investigate users' kill time behavior based on **1000K+** mobile phone-sensor and screenshot data, which is collected by our self-developed Android App via Java and uploaded to Firebase.
- Employed a two-stage clustering approach to separate users into four groups according to the patterns of their phone-usage behaviors, and then built a fusion model for each group, yielding overall strong performance on AUROC.

Knowledge Distillation & K-Means Clustering for Large Generative Models ***[Arxiv] [code]*** *Sept. 2023 – Dec. 2023*

- Guided both CLIP text and image encoders with Llama2 using offline K-means clustering soft labels and knowledge distillation.
- Achieved a **10%** improvement in Exact Match (EM) over the pre-trained CLIP's text encoder on the CC3M dataset and enhanced image attribute classification accuracy by **5%** with the proposed image encoder on the AWA2 dataset.

Model Compression for Real Time Object Detection ***[ICMR'21 (competition)] [code]*** *July 2020 – June 2021*

- Applied structural pruning, knowledge distillation and quantization on YOLOv4. The developed models not only fit for embedded systems (Ex: NVIDIA Jetson TX2) but also achieve higher FPS and mAP at the same time on the multi-spectral infrared dataset.
- Winner of "2021 ACM ICMR Embedded Deep Learning Object Detection Model Compression Competition for Traffic in Asian Countries".

SKILLS

- **Programming:** C/C++, Python (Package: PyTorch, Tensorflow, PySpark, PyTest), Java, SQL, Shell Script, Scala (Spark), MATLAB, R
- **DevOps & Tools:** GCP, Docker, K8s, Git, Jenkins, Argo CD, Helm, Prometheus, Airflow, Grafana, Clickhouse, InfluxDB, MySQL