Tsung-Huan Yang

jason101805@gmail.com | huan80805.github.io

EDUCATION

National Taiwan University (NTU) Taipei, Taiwan
Bachelor of Science in Electrical Engineering Sep 2018 – Jun 2022

- Overall GPA: 3.62/4.30

PUBLICATIONS

[1] Tzuching Lin*, **Tsung-Huan Yang***, Ko-Wei Huang, Hong-Yu Li, I-Bin Liao, Yung-Hui Li, and Lun-Wei Ku. "Learning to Elicit LLM Personality Traits", under review in *ACL ARR 2024*. (*equal contribution)

[2] **Tsung-Huan Yang**, Ko-Wei Huang, Yung-Hui Li, and Lun-Wei Ku. "Preserving Safety in Fine-Tuned Large Language Models: A Systematic Evaluation and Mitigation Strategy", *NeurIPS Safe Generative AI Workshop*, 2024

[3] Tzu-Quan Lin, **Tsung-Huan Yang**, Chun-Yao Chang, Kuang-Ming Chen, Tzu-hsun Feng, Hung-yi Lee, and Hao Tang. "Compressing Transformer-based Self-supervised Models for Speech Processing", *arXiv preprint arXiv:2211.09949*, 2022.

RESEARCH EXPERIENCE

Natural Language Processing and Sentiment Analysis Lab, Academia Sinica Taipei, Taiwan
Research Assistant | Advisor: Dr. Lun-Wei Ku Apr 2023 – Now

- Topic: LLM Personality Control [1]
  - Designed an automated prompt engineering pipeline leveraging GPT-4 to generate 6,000 personality-labeled training samples for BERT fine-tuning in personality classification
  - Implemented soft Q-learning to optimize personality-evoking tokens, validated with psychological questionnaires
- Topic: Defending LLM Fine-tuning Attacks [2]
  - Benchmarked three detoxification methods (supervised fine-tuning, proximal policy optimization, and direct preference optimization) against safety degradation, quantifying their impact on alignment preservation
  - Conducted subspace similarity analysis to identify effective strategies to defend fine-tuning attacks

Eighth Frederick Jelinek Memorial Summer Workshop, Johns Hopkins University Baltimore, MD
Member of Efficient Speech Processing Group | Advisor: Prof. Hao Tang Jun 2022 – Aug 2022

- Topic: Speech Model Compression [3]
  - Implemented iterative attention head pruning and proposed head skipping and 1D convolution swapping strategies
  - Conducted systematic analysis of compression-performance tradeoffs across model sparsity, downstream task performance (ASR, emotion recognition), and computational efficiency

Speech Processing and Machine Learning Lab, NTU Taipei, Taiwan
Undergraduate Research Student | Advisor: Prof. Hung-yi Lee Mar 2021 – Jan 2022

- Topic: Speech Accent Transfer
  - Fine-tuned StarGAN-VC to transfer accent with Mel-cepstral distortion < 10 RMSE, disentangling accent features

WORK EXPERIENCE

Institute of Information Science, Academia Sinica Taipei, Taiwan
Research Assistant Apr 2023 – Now

- Taiwan Visual-Language-Model 70B Project
  - Orchestrated a large-scale data pipeline processing 50M image-text-paired online and private documents
  - Aligned ViT visual encodings with Llama3-Taiwan-LLM-70B, utilizing tensor parallelism to perform distributed training on NVIDIA Taipei-1 computing cluster

- GenAI Safety Assessment Project, funded by *National Institute of Cyber Security*
  - Constructed a dataset, featuring 300,000 human-AI dialogue entries and 100,000 comments crawled from Taiwanese online forums, for building a culturally-aware safeguard model for Taiwanese LLMs
  - Developed automatic red-teaming methods, including black-box (RL-based) and white-box (adversarial suffix searching) attacks, to probe LLM vulnerabilities
- Managed and maintained servers across local hardware resources and Microsoft Azure cloud computing services
- Regularly presented technical knowledge on up-to-date research to 30 international colleagues and research fellows

Applied Materials Hsinchu, Taiwan
Computer Vision Engineering Intern Mar 2022 – Sep 2022

- Defect Detection
  - Fine-tuned YOLOv7 to detect defects on manufactured eyepieces, achieving F1 score >0.95 on all defect types
  - Devised innovative data augmentations to transfer models to unseen defect types with only 20 training samples

TEACHING EXPERIENCE

Advanced Natural Language Processing, Chung Gung University New Taipei, Taiwan
Laboratory Instructor Mar 2024, Oct 2024

Large Language Model, ChungHwa Telecom Training Institute New Taipei, Taiwan
Teaching Assistant Feb 2024 – Sep 2024

Natural Language Processing, MediaTek Inc. Hsinchu, Taiwan
Laboratory Instructor Jul 2024 – Aug 2024

SELECTED PROJECTS

Speech Bot
Network and Multimedia Course Final Project Mar 2022 - Jun 2022

- Distilled a lightweight automatic speech recognition model and employed AI conversation cloud service
- Utilized TensorRT to deploy models on Jetson Nano with 1.4x speedup and 64% memory reduction

Skull Fracture Detection
Deep Learning for Computer Vision Course Final Project Nov 2021 - Jan 2022

- Optimized FasterRCNN model for medical image fracture using targeted ROI cropping techniques, achieving F1>0.6 and ranking in top 10% of 120 competitors

Club Management and Activity Subscription Website
Web Programming Course Final Project Oct 2021 - Dec 2021

- Developed real-time notification system with React, implementing efficient broadcast management
- Integrated GraphQL for optimized data fetching and selective content delivery to reduce network overhead

SKILLS

Programming Language: Python, C, C++, Java, JavaScript, Go, TeX

Infrastructure & Platform: Azure, Slurm, GCP, MongoDB

ML/NLP Library: PyTorch, TensorFlow, Scikit-learn, Megatron, Transformers, TRL, LangChain, LlamaIndex

EXTRACURRICULAR ACTIVITIES

Sunny Coconut Social Service Club, NTU Taipei, Taiwan
Camp Organizer Jan 2020 – Jan 2021

- Led 20 volunteers to organize educational camps serving 50+ elementary students from low-income backgrounds
- Designed STEM-based curriculum, coordinated staff, and led review meetings to optimize future activities