

Sin-Han Yang

✉ harry900302@gmail.com | 🏠 sinhanyang.github.io | 🗣️ [SinHanYang](#) | 💼 [sinhanyang](#) | 🌐 [Sin-Han Yang](#)

Education

National Taiwan University (NTU)

B.S. in Computer Science and Information Engineering

- Double Major in Physics

Taipei, Taiwan

Sep. 2019 - Jun. 2024

University of Birmingham

Exchange Student in Computer Science

Birmingham, UK

Sep. 2022 - Jan. 2023

Research Interests

Understanding deep learning as well as improving its robustness and generalization, both from theoretical and empirical perspectives.

Papers

Conference Paper

[C1] **Sin-Han Yang**, Chung-Chi Chen, Hen-Hsen Huang and Hsin-Hsi Chen, [Entity-Aware Dual Co-Attention Network for Fake News Detection](#), *Findings of the Association for Computational Linguistics: EACL*, 2023.

Journal Paper

[J1] **Sin-Han Yang**, Tuomas Oikarinen, Tsui-Wei Weng, [Concept-Driven Continual Learning](#), *Transactions on Machine Learning Research (TMLR)*, 2024.

Work Experience

RIKEN Center for Advanced Intelligence Project

Research Assistant

Tokyo, Japan

Aug. 2024 - PRESENT

- Advisor: [Dr. Emtiyaz Khan](#)
- **Project 1** : Weight Perturbation implicitly induces label noise.
- Derived the implicit label noise from Variational Learning, which is learned for each sample.
- Empirically showed that Variational Learning outperforms label smoothing and is comparable with SAM.
- **Project 2** : Use Bayesian Learning and model merging principles to improve continual learning.
- Theoretically showed that reducing gradient mismatch can ideally achieve batch training's accuracy.
- Proposed that selecting memory based on Hessian can efficiently reduce gradient mismatch.

Research Experiences

Computer Science and Engineering Department, UC San Diego

Visiting Student

Remote

Jun. 2022 - Aug. 2024

- Advisor: [Prof. Tsui-Wei \(Lily\) Weng](#)
- **Project 1** : LLM Jailbreak Defense with formal guarantee.
- Applied Random Smoothing on target LLMs, and derived the corresponding robustness certification.
- Defended major jailbreak algorithms, which reduces attack success rate by up to 78%.
- **Project 2** : Use model's interpretability to improve performance in continual learning [J1]
- Controlled interpretable neurons to understand the continual learning process and migrate the forgetting.
- Proposed new methods that are comparable with previous works, but significantly boost the interpretability.

Nature Language Processing Laboratory, NTU

Undergraduate Researcher

Taipei, Taiwan

Nov. 2021 - Feb. 2024

- Advisor: [Prof. Hsin-Hsi Chen](#)
- Focused on fake news detection, design a new attention-based architecture for interpretability **[C1]**
- The new architecture outperforms baselines in standard benchmarks.
- Used model's interpretability to analyze the key words and sentences for final predictions.

Electrical and Computer Engineering Department, Princeton University

Visiting Student

Remote

Jun. 2023 - Sep. 2023

- Advisor: [Prof. Jason D. Lee](#)
- Worked on the theoretical aspect of continual learning from representation learning. [\[Working Note\]](#)
- Showed that in task incremental learning, models can learn nonlinear representations with bounded errors.
- Challenged the class incremental learning, and point out few-shot continual learning as a future direction.

Research Center for Information Technology Innovation, Academia Sinica

Research Assistant

Taipei, Taiwan

Jan. 2021 - Jan. 2022

- Advisor: [Dr. Gen-Cher Lee](#)
- Modified communication software's source code to extend functionality.
- Gained the ability to understand, modify and test big open source software.

Honors & Awards

Appier Best R&D Award and CSIE 3rd Place NTU CSIE Bachelor Research Exhibition

Taipei, Taiwan

2023

College Student Research Scholarship National Science Council (NSC)

Taipei, Taiwan

- NSC scholarship for excellent students based on written research proposal

2023

NTU Y.L.LIN Scholarship

Taipei, Taiwan

- For exchange students

2022

NTU Dean's List Award

Taipei, Taiwan

2021

Final Selection and Training Camp IPHO (International Physics Olympics) Taiwan team

Taipei, Taiwan

2018

Professional Activities

- **Reviewer** ICLR 2025
IEEE Transactions on Knowledge and Data Engineering
NeurIPS 2024 Safe Generative AI Workshop

Skills

- **Programming** Python, MATLAB, C/C++
- **Others** PyTorch, Git, \LaTeX