ZHEN ZHANG

+1(805)951-7538 | email: zhen_zhang@ucsb.edu | Homepage

EDUCATION

UC Santa Barbara — California, United States
Doctor of Philosophy in Computer Science — Sept 2023 – Expected July 2028

- **Advisor:** Shiyu Chang

Tsinghua University — Beijing, China
Bachelor of Science in Mathematics and Physics — Sept 2019 – Expected July 2023

- **Honors & Awards:** Silver Medal in the 35th Chinese Physics Olympiad, Academic Excellence Award (2021-2022)
- **Relevant Courses:** Foundation of Object-Oriented Programming, Data Structures, Discrete Mathematics, Theory of Computer Network, Artificial Neural Networks

PUBLICATIONS

Certified Robustness for Large Language Models with Self-Denoising (preprint)
Zhen Zhang, Guanhua Zhang, Bairu Hou, Wenqi Fan, Qing Li, Sijia Liu, Yang Zhang, Shiyu Chang

Parameter-Efficient Cross-lingual Transfer Learning of Vision and Language Models via Translation-based Alignment (preprint)
Zhen Zhang, Jialu Wang and Xin Eric Wang

Sparse Structure Search for Delta Tuning (NeurIPS 2022)
Shengding Hu*, **Zhen Zhang***, Ning Ding, Yadao Wang, Yasheng Wang, Zhiyuan Liu, Maosong Sun
(* indicates equal contribution)

RESEARCH EXPERIENCE

University of California, Santa Barbara — Online
Ph.D. Student of Professor Shiyu Chang — Jul 2022 – Present
Shiyu's Lab

- Proposed SelfDenoise, a new method to certify robustness of large language models against input perturbations. It uses the model's own ability to denoise corrupted inputs in a self-supervised way, improving certified robustness under randomized smoothing. The method achieves higher robustness than prior approaches without extra training or model access.

Tsinghua University — Beijing, China
Research Assistant to Professor Zhiyuan Liu — Jan 2022 – Present
THUNLP

- Undergraduate Thesis
- Implement Reinforcement Learning from Human Feedback (RLHF) on a bilingual LLM CPM-Bee.
- Empowering models with conversational, multitasking capabilities while making their responses useful and harmless
- Ranked 11th on the ZeroCLUE Zero-Shot Learning Leaderboard.

University of California, Santa Cruz — Online
Summer Internship with Professor Xin Eric Wang — Jul 2022 – Present
Eric Lab

- Addressed the multilingual disparity of vision and language foundation models such as CLIP.
- Proposed a unified framework by translation and taking advantage of parallel data. Methods for mapping multilingual texts to the same language space and aligning them improve multilingual disparity.

Tsinghua University — Beijing, China
Research Assistant to Professor Zhiyuan Liu — Jan 2022 – Present

THUNLP

- Participated in developing a toolkit called OpenDelta, an open-source framework for parameter-efficient tuning.
- By using this toolkit, users could easily implement various types of parameter-efficient tuning with preferred pre-trained models.
- Tested and adapted more Transformer-based models to OpenDelta.
- This work has been accepted to ACL 2023 Demo track.

Tsinghua University Beijing, China
Research Assistant to Professor Zhiyuan Liu Jan 2022 – May 2022
THUNLP

- Presented a method that automatically Search for the Sparse Structure of Delta Tuning (S^3Delta).
- S^3Delta conducts the differentiable parameter-efficient tuning (PET) structure search through bi-level optimization, and uses Shifted Global Sigmoid method to explicitly control the number of trainable parameters and optimize the combination of various current PET methods.
- S^3Delta preserves more than 99% of fine-tuning performance with 0.01% trainable parameters.
- This work has been accepted by NeurIPS 2022.

SELECTED COURSE PROJECTS

Paper Reproduction

- Course project for course Artificial Neural Networks.
- This repository provides a reproduction based on the framework Jittor for the paper 'Table-to-text Generation by Structure-aware Seq2seq Learning' selected in AAAI2018.

Slow Electron Velocity Imaging

- Course project for course Big Data in Experimental Physics taught by Benda Xu.
- Trained ResNet34 from scratch to accurately reconstructs the coefficients of the Lejeune equation in the electron velocity field.