Last Updated on June 26, 2024

Chris Yuhao Liu

✉ yliu298@ucsc.edu   chrisliu298   🏠 chrisliu298.ai   Google Scholar

RESEARCH INTERESTS

My research focuses on machine unlearning and alignment for large language models. I'm particularly interested in fine-grained control of large language models' behaviors post-training. Previously, I worked on fundamental problems in deep learning, including double descent, regularization, and data scaling laws.

EDUCATION

Ph.D. in Computer Science and Engineering, University of California, Santa Cruz — 2023 - present
Advisors: Yang Liu and Jeffrey Flanigan

M.S. in Computer Science and Engineering, University of California, Santa Cruz — 2021 - 2023
Thesis: Understanding the Role of Optimization and Loss Function in Double Descent
Advisor: Jeffrey Flanigan
GPA: 4.0

B.S. in Computer Sciences, University of California, Santa Cruz — 2017 - 2021
Advisor: Jeffrey Flanigan
Honor: Honors in the Major, Cum Laude
GPA: 3.71

RESEARCH EXPERIENCE

Graduate Student Researcher — 2023 - present
REAL, University of California, Santa Cruz — Santa Cruz, CA, USA

Research Intern — 2022
REAL, University of California, Santa Cruz — Santa Cruz, CA, USA

Student Researcher — 2020 - 2023
JLab, University of California, Santa Cruz — Santa Cruz, CA, USA

PUBLICATIONS AND PREPRINTS

Large Language Model Unlearning via Embedding-Corrupted Prompts [Paper] [Project page]
Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, Yang Liu
In Submission

Advancing Machine Unlearning Evaluation Requires Rethinking Retraining
Chris Yuhao Liu, Zonglin Di, Jeffrey Flanigan, Yang Liu
In Submission

Understanding the Role of Optimization in Double Descent [Paper] [Poster]
Chris Yuhao Liu and Jeffrey Flanigan
NeurIPS 2023 Workshop on Optimization for Machine Learning

Understanding the Role of Optimization and Loss Function in Double Descent [Paper]
Chris Yuhao Liu
Master Thesis

Toward Disentangling Double Descent and Information Flow in Deep Neural Networks [Paper], [Code]
Chris Yuhao Liu, Brendan King, Jing Gu

Learning to Extract Compact Vector Representations from Weight Matrices [Paper], [Code], [Slides]
Chris Yuhao Liu, Zichao Li

Sample Complexity Scaling Laws For Adversarial Training [Paper], [Code]
Chris Yuhao Liu

PROJECTS

Awesome Representation Engineering [GitHub]
A comprehensive list of work on representation engineering and activation steering.

Awesome Large Language Model Unlearning [GitHub]
A comprehensive list of work on machine unlearning in large language models.

Structural Risk Minimization for Deep Neural Networks
A new regularization method based on structural risk minimization that directly minimizes the generalization gap.

What Determines Sample Complexity Rate in Practice?
We empirically estimate the power-law exponents of various model architectures and study how they are altered by a wide range of training conditions for classification.

Faster Sample Complexity Rates With Ensemble Filtering
We present a dataset filtering approach that uses sets of classifiers, similar to ensembling, to estimate noisy (or non-realizable) examples and exclude them so a faster sample complexity rate is achievable in practice.

Conditional Research Paper Abstract Generation [Code]
A GPT-2 that generates paper abstact based on the given title, trained on all cs.AI, cs.LG, cs.CL, and cs.CV papers on arXiv. This was the winner of the Generative Modeling Competition for CSE 142 in Spring 2020 at UC Santa Cruz.

TAPT: Text Augmentation Using Pre-Trained Transformers With Reinforcement Learning [Code]
A classification data generator trained using PPO.

Sentiment Analysis with Transformers [Code]
A RoBERTa sentiment classifier. This was the winner of the Sentiment Analysis Competition of CSE 142 in Spring 2020 at UC Santa Cruz.

HONORS AND FELLOWSHIPS

2023	**Regents Fellowship**, University of California, Santa Cruz	Santa Cruz, CA, USA
2023	**Department Fellowship**, University of California, Santa Cruz	Santa Cruz, CA, USA
2021	**Honors in the Major, Cum Laude**, University of California, Santa Cruz	Santa Cruz, CA, USA
2021	**Dean's Honors**, University of California, Santa Cruz	Santa Cruz, CA, USA
2020	**Dean's Honors**, University of California, Santa Cruz	Santa Cruz, CA, USA
2019	**Dean's Honors**, University of California, Santa Cruz	Santa Cruz, CA, USA
2018	**Dean's Honors**, University of California, Santa Cruz	Santa Cruz, CA, USA
2017	**Dean's Honors**, University of California, Santa Cruz	Santa Cruz, CA, USA

TEACHING EXPERIENCE

Teaching Assistant, *University of California, Santa Cruz*, Santa Cruz, CA, USA 2021 - present

- CSE 20 Introduction to Python (Fall 2021, Spring 2022, Fall 2022, Winter 2024)
- CSE 30 Programming Abstractions: Python (Spring 2023)
- CSE 40 Machine Learning Basics (Spring 2024)
- CSE 144 Applied Machine Learning (Winter 2022)

Tutor and Reader, *University of California, Santa Cruz*, Santa Cruz, CA, USA 2020

- CSE 142 Machine Learning (Fall 2020)

SERVICE

Volunteer, International Conference on Machine Learning 2021
Volunteer, International Conference on Learning Representations 2021

SKILLS

Knowledge areas: Machine learning, deep learning, natural language processing
Machine Learning frameworks: PyTorch, Hugging Face Transformers, DeepSpeed, scikit-learn, Keras
Programming languages: Python, R, Shell
Data analysis and visualization: Pandas, NumPy, Matplotlib, Seaborn, Weight & Biases
Writing: LaTeX, Markdown
Miscellaneous: Git, Kubernetes