Yilun Kuang

yk2516@nyu.edu | github.com/YilunKuang | linkedin.com

Research Interests

My research interests lie in the intersection of theoretical neuroscience and deep learning, which includes

- Self-Supervised Representation Learning, Multi-Modal Machine Learning, Language Modeling
- Representational Geometry, Neural Tangent Kernal and Mean Field Perspective of Training Dynamics

Education

New York University cumulative GPA: 3.85/4.00

B.A Honors Mathematics & Computer Science *Sep. 2020 – May 2023*

- Relevant coursework: Cloud and Machine Learning (Grad), NLP with Representation Learning (Grad), Math Tools for Data Science (Grad), Mathematics of Deep Learning (Grad), Foundations of Machine Learning (Grad), Honors Numerical Analysis, Honors Analysis I & II, Parallel Computing, Object-Oriented Programming

Pitzer College cumulative GPA: 3.9/4.0

B.A Cognitive Science; B.A. Philosophy *Sep. 2019 – May 2020*

- Relevant coursework: Intro to Cognitive Science, Foundations of Neuroscience, Topics in Neurophilosophy

Research Experience

Flatiron Institute (Simons Foundations) Jan 2022 – Present

Guest Researcher & Summer Research Intern (Advisor: Prof. SueYeon Chung) *New York, NY*

- Guest Researcher (Fall 2022)
- Summer Research Intern (Summer 2022)
  - Participate in developing a self-supervised representation learning algorithm based on the Principle of Maximal Manifold Capacity drawn from statistical physics literatures of perceptron/manifold capacity.
  - Perform Distributed Training of ImageNet using 32 A100 GPUs. The linear evaluation accuracy of our model is on par with state-of-the-art self-supervised learning method like SimCLR and Barlow-Twins.
  - Perform SSL training on the CIFAR10 / ImageNet dataset using PyTorch Distributed Data Parallel (DDP) framework with 32 NVIDIA A100-SXM4-40GB GPUs in a multi-node setting with NCCL backend.
  - Apply composer, FFCV, Albumentations, Pillow-SIMD and other packages for training and inference efficiency optimization. Reduce training time cost from 8 days to 1 days
  - The linear evaluation accuracy of our model is on par with state-of-the-art self-supervised learning method like SimCLR and Barlow-Twins. Manuscript in preparation for ICLR 2023 Submission.
- Research Assistant (Spring 2022)
  - Implement Mean-Field Theoretic Manifold Analysis (MFTFA) on the brain-inpsired VOneNet models and evaluate the contribution of biologically realistic divisive normalization modules to the model's adversarial robustness under projected gradient descent attack.

NYU Machine Learning for Language Group June 2021 – May 2022

Natural Language Processing Research Assistant (Advisor: Prof. He He) *New York, NY*

- Fine-tune BART and GPT2 on In-Distribution datasets (ID) and implement ROUGE, PPL, AUROC evaluation metrics on several Out-of-Distribution (OOD) text summarization datasets on NYU HPC Greene clusters.
- Run experiments on the Singularity Containers and use the Slurm resource allocation system for large-scale model pretraining, fine-tuning, and evaluations. Monitor the hardware and model performance with customized NVIDIA CLI interface. Hands-on model checkpoints, loggings, and debugging.
- Follow up on literature on out-of-distribution detection, prompt tuning, adapters, and develop a new algorithm for unsupervised domain adaptation in text summarization.

NYU Shanghai Speech and Language Neuroscience Group Oct 2020 – May 2021

Undergraduate Research Assistant (Advisor: Xing Tian) *Shanghai, China*

- Designed a Bayesian hierarchical model for self-other distinction of auditory stimuli. Developed object-oriented codes for the psycho-physical experiment.
- Attend weekly research talk in the lab, read on research papers, derived a Bayesian hierarchical generative model to characterize behavioral responses in speech perception.

Project Experience

SiFT Algorithm for Improved Generalization Feb 2021 – May 2021

- Course Project in DS-UA 203 Machine Learning for Language Understanding. Implemented Scale-invariant-Fine-Tuning (SiFT) algorithm and the Smoothness-Inducing Adversarial Regularization and Bergman Proximal Point Optimization (SMART) algorithm for virtual adversarial training on BERT and DeBERTa and achieved better adversarial robustness and model generalization.
- Implemented a novel model generalization benchmark: Perform fine-tuning for BERT and DeBERTa on In-Domain Twitter Hate Speech dataset and fine-tune on 10 percentage of Out-of-Domain UCI Sentiment dataset and evaluate the testing result.

Scalable MNIST Inference on IBM Cloud Kubernetes Nov 2021 – Dec 2021

- Course Project in CSCI-GA 3033 Cloud and Machine Learning. Develop Cloud Native dockerized MNIST hand-written digit online inference microservice with Python Flask based on IBM Cloud Kubernetes CPU Cluster to replace GPU chips. Achieve 0.78 inference speedup but with 2.28 cheaper business solution for model serving using a CPU cluster.
- Implement multithreading techniques to create multiple simultaneous client connections to the microservice and evaluate the performance bottleneck. Work with yaml file write-up and docker networking.

NTK and Mean Field Limit for Overparametrized Network Training Dynamics Feb 2022 – May 2022

- Course Project in CSCI-GA 3033 Mathematics of Deep Learning. Compared the lazy learning (NTK) vs. active learning (Mean Field Limit) regimes for overparametrized neural networks. Project Report Link
- Numerically simulated two regimes for finite neural networks and compare the training dynamics.

Adversarial Defense via Natural Supervision and Quantization Feb 2022 – May 2022

- Course Project in CSCI-GA 2566 Foundations of Machine Learning. Implement natural supervision method (contrastive loss estimation) and quantization methods for improved adversarial robustness of deep neural network during inference

Poisson-Gamma Neural Variability in the Visual Cortex Oct 2020 – Dec 2020

- Course Project in NEUR-UA 302 Computational Neuroscience. Derived a poisson-gamma probabilistic model to characterize the noise distribution in visual neuron responses.
- Simulated neural population encoding and performed maximum likelihood decoding and statistical tests to validate the poisson-gamma model in Matlab

Presentation

- Maximal Coding Rate Reduction. Aug 5, 2022. Flatiron Institute NeuroAI & Geometry Reading Group
- Bayes Point Machine and Kernels. Aug 26, 2022. Flatiron Institute NeuroAI & Geometry Reading Group

Services

- Grader for DS-GA 1012 Natural Language Understanding and Computational Semantics at NYU (Spring 2022)

Workshops/Seminars

- Neuromatch Academy 2020: Computational Neuroscience

Skills/Other

- Honors and Award: Nicholas and Andrea Ferrara Research Scholar, Dean's List (2020-2022), DURF Grant Recipient
- Student Journal: Paper selected by the NYU Integrative Psychology Review (NYU Undergraduate Psychology Research Journal)
- Competition: Meritorious Winner in 2021 Mathematical Contest in Modeling (MCM) (Top 7%)
- Framework: PyTorch, TensorFlow, Transformers, Scikit-Learn, CUDA, OpenMP, MPI
- Language: C++, C, Python, Java, Matlab, Julia, SQL, Bash Shell