Ruomin Huang

Email: ruomin.huang@duke.edu
Personal Homepage: fockee.github.io

Research Interests

•Transformers and Language Models

– **In-Context Learning; Parametric Memory; Post-Training; SFT; RLVR**

•Algorithmic Aspect of Machine Learning

– **Efficient and Robust Machine Learning**

Education

•Duke University *2023 - present*

Computer Science Ph.D. student, supervised by Prof. Rong Ge

•School of Data Science, University of Science and Technology of China (USTC) *2020 - 2023*

Master's student in Data Science, supervised by Prof.Hu Ding

•School of Mathematical Science, University of Science and Technology of China (USTC) *2016 - 2020*

Bachelor of Science in Information and Computational Science

Papers

•Transformers and Language Models

– **Ruomin Huang**, Eshaan Nichani, Jason D. Lee and Rong Ge, "Fine-Tuning Dynamics of In-Context Factual Recall in Transformers," preprint.
– **Ruomin Huang**, Rong Ge, "Task Descriptors Help Transformers Learn Linear Models In-Context," ICLR 2025.
– Roy Xie, Junlin Wang, **Ruomin Huang**, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, Bhuwan Dhingra, "ReCaLL: Membership Inference via Relative Conditional Log-Likelihoods," EMNLP 2024

•Algorithmic Aspect of Machine Learning

– Wanlin Zhang, Weichen Lin, **Ruomin Huang**, Shihong Song, Hu Ding, "To Tackle Adversarial Transferability: A Novel Ensemble Training Method with Fourier Transformation," ICLR 2025
– Jiaxiang Chen, Weichen Lin, **Ruomin Huang**, Hu Ding, "An Effective Dynamic Gradient Calibration Method for Continual Learning," ICML 2024.
– **Ruomin Huang**, Jiawei Huang, Wenjie Liu, Hu Ding, "Coresets for Wasserstein Distributionally Robust Optimization Problems," NeurIPS 2022 **(spotlight)**.
– Jiaxiang Chen, Qingyuan Yang, **Ruomin Huang**, Hu Ding, "Coresets for Relational Data and The Applications," NeurIPS 2022 **(spotlight)**.
– Jiawei Huang*, **Ruomin Huang***, Wenjie Liu*, Nikolaos M. Freris, Hu Ding, "A Novel Sequential Coreset Method for Gradient Descent Algorithms," ICML 2021 **(spotlight)**.

Selected Research Experiences

•Fine-Tuning Dynamics of In-Context Factual Recall in Transformers *May 2025 - May 2026*

– Designed a two-step factual recall task and fine-tuned a one-layer Transformer to infer a hidden relation from context and retrieve the corresponding fact from a frozen associative memory.
– Proved that perturbed gradient descent escapes a symmetric saddle and learns pairwise attention in a minimal model. Showed that low-temperature fine-tuning needs only polylogarithmically many examples in the number of stored facts.
– Achieved high relation-prediction accuracy with 13 synthetic fine-tuning sequences and reproduced pairwise attention with a pretrained associative memory.

•Emergence of In-Context and In-Weight Learning Circuits in Disentangled Transformers *Sep 2024 - Sep 2025*

– Studied the training dynamics of disentangled transformers in the presence of in-context learning (ICL) data and in-weight learning (IWL) data.
– Theoretically characterized a hyperparameter regime (e.g., noise levels, learning rates, sequence length) that jointly enables ICL and IWL, and a distinct regime in which ICL and IWL compete.
– Empirically discovered various circuits for ICL.

•Task Descriptors Help Transformers Learn Linear Models In-Context *Sep 2023 - Oct 2024*

– Studied the in-context learning of mean-varying linear regression problem in the presence of task descriptors.
– Proved a global convergence result for Transformers trained with task descriptors.

– Gave an interpretation on how Transformers use task descriptors to learn in-context.
– Empirically found a clear pattern in the parameters of well-trained Transformers which matches our theory.

•Large-Scale Wasserstein Distributionally Robust Optimization *Mar 2022 - Aug 2022*

– Studied machine learning problems with ambiguous training data, especially the Wasserstein distributionally robust optimization (WDRO) problems.
– Proposed the first dual coreset method to accomplish the data compression for WDRO problems and theoretically analyzed the performance of the algorithm.
– Evaluations showed that our method had a lower worst-case risk compared with baseline methods.

•Data Sampling for Empirical Risk Minimization *Nov 2020 - May 2021*

– Extended the layer-sampling algorithm for clustering problems to general ERM problems so as to construct coresets.
– Theoretically analyzed the sampling complexity of the algorithm.
– Evaluations showed that our method outperformed baseline methods on real-world benchmarks.

Skills

• Python, PyTorch, NumPy, Git, Linux, Weights & Biases, Codex