STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. DISSERTATION DEFENSE

Candidate: Wuxinlin Cheng
Degree: Doctor of Philosophy
School/Department: Charles V. Schaefer, Jr. School of Engineering and Science/ Electrical and Computer Engineering
Date: Wednesday, April 9th, 2025
Time/Location: 11:00 a.m. / Burchard 219
Title: Stability Analysis of Machine Learning Models on the Manifolds

Chairperson: Dr. Zhuo Feng, Department of Electrical and Computer Engineering, School of Engineering & Sciences

Committee Members: Dr. Shucheng Yu, Department of Electrical and Computer Engineering, School of Engineering & Sciences
Dr. Koduvayur Subbalakshmi, Department of Electrical and Computer Engineering, School of Engineering & Sciences
Dr. Yue Ning, Department of Computer Science, School of Engineering & Sciences
Dr. Tian Han, Department of Computer Science, School of Engineering & Sciences

ABSTRACT

Recent advances in machine learning have yielded remarkable gains across vision, language, and graph-based tasks. Yet, persistent vulnerabilities to adversarial perturbations raise significant challenges for deploying these systems in safety-critical domains. This dissertation tackles the robustness problem from a unified, *spectral* perspective on manifold learning. It develops near-linear-time frameworks that analyze and enhance stability by studying how input–output distance mappings become distorted under learned models.
First, **SPADE** (Spectral Method for Black-Box Adversarial Robustness Evaluation) addresses general deep neural networks. SPADE constructs low-dimensional input and output manifolds and uses effective-resistance distances to detect how small input changes can trigger large output distortions. This yields an upper bound on the model's Lipschitz constant without white-box model access. Empirical tests on MNIST and CIFAR-10 confirm that SPADE pinpoints highly fragile data samples and strengthens adversarial training.

Next, **SAGMAN** (Stability Analysis of Graph Neural Networks on the Manifolds) focuses on the growing role of GNNs. While GNNs excel at modeling structured data, modest edge or feature modifications can lead to drastic performance drops. SAGMAN's key step is a scalable Graph Dimensionality Reduction that preserves spectral properties, enabling it to quantify each node's vulnerability. This insight not only guides more targeted adversarial attacks but also underpins a strategy of judiciously "reinserting" edges to fortify GNN stability on citation, recommendation, and large-scale benchmark graphs.

Finally, **SALMAN** (Stability Analysis of Language Models Through the Maps Between Graph-based Manifolds) extends these ideas to Transformer-based language models, from smaller BERT-like architectures to large-scale GPT and Llama systems. By converting token-level outputs into continuous embeddings and learning probabilistic graphical models, SALMAN identifies per-sample fragility under text edits. Ranking

by fragility yields more efficient adversarial prompt attacks, while selectively up-weighting non-robust data during fine-tuning preserves pre-trained representations and boosts overall resilience.

Together, these methods illustrate how spectral analysis of manifold distortions can unify adversarial robustness studies in a scalable manner. By identifying and correcting the most vulnerable data regions, the proposed frameworks provide principled, model-agnostic approaches that strengthen deep learning systems against adversarial disruptions—across standard vision benchmarks, graph domains, and next-generation language models.