Proposal Defense
Doctor of Philosophy in Computer Science

“Towards Efficient and Robust Deployment of Graph Deep Learning” by Yue Dai

Date: May 6, 2024
Time: 11:00 a.m. – 12:30 p.m.
Place: 6329 Sennott Square, 3810 Forbes Ave, Pittsburgh, PA 15213

Committee:
- Youtao Zhang, Professor, Department of Computer Science, School of Computing and Information
- Xulong Tang, Assistant Professor, Department of Computer Science, School of Computing and Information
- Stephen Lee, Assistant Professor, Department of Computer Science, School of Computing and Information
- Jun Yang, Professor, Department of Electrical and Computer Engineering, Swanson School of Engineering

Abstract:
Inspired by the success of Graph Neural Networks (GNNs), recent graph deep learning studies have introduced GNN-based models like Graph Matching Networks (GMNs) and Temporal Graph Neural Networks (TGNNs) for diverse tasks in various domains such as social media, chemistry, and cybersecurity. Despite their success in improving prediction accuracy, deploying these models efficiently and robustly in real-world settings presents considerable challenges. The reasons are three-fold: First and foremost, these GNN-based deep learning models suffer from suboptimal inference latencies that can hardly fulfill the requirements of the real-world applications. Second, the training efficiency and scalability of the GNN-based models is limited, making it hard to efficiently and effectively develop and prepare the model for the targeted applications. Beyond these challenges, GNN-based models are limited by their robustness towards adversarial attacks, making them hard to be deployed under the real-world security and privacy concerns.

My research conducts full-stack optimizations on the GNN-based deep learning models. First, my work focuses on accelerating inferences by proposing a segmented quantization method for GNNs, designing a redundancy-free accelerator and a flexible GPU runtime framework for GMNs. Second, my work focuses on optimizing the training efficiency and accuracy of these models by proposing a dependency aware TGNN training framework, a general layer-freezing approach for model training, and designing a novel GMN model that adaptive to various graph inputs. Lastly, I investigate potential adversarial attacks and defenses on TGNNs under real-world setting.