COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Salar Latifi

Efficient and Dependable Deep Learning Systems

Wednesday, December 7, 2022
1:00 – 3:00pm
Virtual – Zoom

ABSTRACT: Deep Learning (DL) is overhauling a plethora of applications including voice assistants, autonomous vehicles and driving assist technologies, and e-commerce. With such a huge impact on human daily life, researchers are pushing towards designing deep learning systems with higher quality and better performance. However, the trend of training bigger and higher quality models, could lead to higher demands of the hardware resources to be able to process and service the daily applications in an efficient and timely manner. On the other hand, Deep DL applications are getting widely deployed in mission critical applications including autonomous vehicles and precision medicine. Therefore, another important question to answer is how to assure the dependability of DL and trustworthiness of their predictions.

In this thesis, I am aiming to propose solutions for both of these problems. The goal is to optimize different deep learning applications with respect to the reliability of their predictions, and improve their inference performance by reducing their latency and energy requirements. First, I will be focusing on vision applications, including Image Classification and Object Detection, for their efficient and safe deployment. I will explore the inherent reliability problems of the off-the-shelf baselines, and how we can end-up with frequent misclassified and undetected objects with high confidence levels. Then, I will discuss how we can improve the reliability of predictions by introducing modular redundancy and test-time augmentation into the system, and by incorporating more intelligent decision making policies.

Finally, I will be focusing on designing hardware-aware and efficient Transformer architectures for the language modeling tasks. First, I will discuss the significant performance costs of the multi-head attention layers. Next, I will introduce the PLANER optimizer which takes an existing Transformer-based network and a user-defined latency target and automatically produces an optimized, sparsely-activated version of the original network that tries to meet the latency target requirements while maintaining baseline accuracy.

CHAIR: Prof. Scott Mahlke