NJIT | INSTITUTE FOR DATA SCIENCE

Data Science Seminar Series

In Collaboration with
The Department of Data Science

Resource-Efficient Machine Learning

Hosted by Hai Phan

Xiaotian Han

Texas A&M University

Date: **Wednesday, February 14, 2024**
Time: **10:30 AM - 11:30 AM (Coffee served at 10:15 AM)**
Location: **GITC Building Room 3600 (3rd floor Seminar Room)**
Web Link: **Zoom Meeting Link**

IIn this talk, I will present my research on resource-efficient machine learning techniques for graph neural networks (GNNs) and large language models (LLMs). These techniques aim to reduce the computational resources required by these models, making them more practical for real-world applications. i) I will discuss my work on accelerating the training and inference of GNNs by connecting them with multilayer perceptrons (MLPs). GNNs can be computationally expensive to train and use due to the large size of graphs. I discovered that the weights between the GNN and MLP layers are transferable. This means that I can pre-train an MLP on node feature only and then use the weights of this pre-trained MLP to initialize the GNN. This significantly reduces the training time of the GNN. Additionally, I creatively reformulated GNNs into a form of Mixup that can be implemented through an MLP, further improving efficiency. ii) I will discuss my work on reducing the computational cost of pretraining LLMs and expanding their context abilities. A major obstacle with LLMs is their immense parameter size, causing slow training/fine-tuning. To accelerate LLMs/pretraining, I proposed GrowthLength, which gradually increases the training sequence. Furthermore, fine-tuning large models for long context understanding can be expensive. To reduce this cost, I developed SelfExtend to extend the context window without requiring costly fine-tuning by constructing a bi-level attention mechanism, which unlocks LLMs' inherent capabilities to handle long contexts.
My research aims to address computational and efficiency challenges in large models, thereby enabling their use in more real-world applications. I hope to democratize state-of-the-art AI capabilities for high-impact societal use cases even with limited resources.

Xiaotian (Max) Han is a Ph.D. candidate in computer science at Texas A&M University, advised by Dr. Xia (Ben) Hu. His research interests lie in artificial intelligence, machine learning, and data science, with a focus on designing resource-efficient deep learning methods for computationally restricted environments. He aims to democratize cutting-edge machine learning for high-impact societal applications with limited resources. He has published over 20 papers, including 11 first-authored papers, in top-tier machine learning conferences and journals such as ICML, ICLR, TMLR, WWW, KDD, IJCAI, AAAI, TKDE, etc. He has also served as reviewer for these top conferences. He was the recipient of the Outstanding Paper Award at ICML 2022 as first author.