Accelerating Sparse Deep Neural Networks using GPUs

Project Investigator (PI): Vishwesh Jatala

Problem Statement:

Deep Neural Network (DNN) has become a popular tool for solving complex problems in various domains, such as social networks, computer vision, natural language processing, and health care. As a result, improving the performance of the DNN applications has been an active area of research for the past few years. Moreover, researchers have been exploring accelerators, like Graphics Processing Units (GPU), to achieve high throughput and bandwidth. However, real-world deep neural networks have grown beyond the memory size of the GPUs. Consequently, the deep learning community is exploiting techniques to sparsify the DNN.

Sparse DNNs, due to irregular memory access patterns, suffer from performance and scalability issues. To address these challenges, this project aims to develop efficient techniques for sparse deep neural networks (DNN).

References:

1. *Sparse Deep Neural Network Graph Challenge*, Jeremy Kepner, Simon Alford, Vijay Gadepally, Michael Jones, Lauren Milechin, Ryan Robinett, Sid Samsi, IEEE HPEC, 2019
2. *Fast Sparse Deep Neural Network Inference with Flexible SpMM Optimization Space Exploration* - Jie Xin, Xianqi Ye, Long Zheng, Qinggang Wang, Yu Huang, Pengcheng Yao, Linchen Yu, Xiaofei Liao, Hai Jin (Huazhong University of Science and Technology)
3. *At-Scale Sparse Deep Neural Network Inference With Efficient GPU Implementation* - Mert Hidayetoglu, Carl Pearson, Vikram Sharma Mailthody (UIUC), Eiman Ebrahimi (Nvidia), Jinjun Xiong (IBM)), Rakesh Nagi, Wen-mei W. Hwu (UIUC)
4. *Studying the Effects of Hashing of Sparse Deep Neural Networks on Data and Model Parallelism* - Mohammad Hasanzadeh Mofrad, Rami Melhem (Univ of Pittsburgh), Yousuf Ahmad, Mohammad Hammoud (CMU Qatar)
5. *A GPU Implementation of the Sparse Deep Neural Network Graph Challenge* - Mauro Bisson, Massimiliano Fatica (NVIDIA)