INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2021
Technology, Social Impact

Generating Efficient Block Sparse Neural Networks using RBGP Framework

ABSTRACT

Sparsity is an essential tool for generating compute and memory efficient neural networks. However current hardware like GPU's can only exploit structured sparsity patterns for better efficiency.
In this work, we propose RBGP(Ramanujan Bipartite Graph Product) framework for generating structured multi level block sparse neural networks by using the theory of Graph products. We also propose to use products of Ramanujan graphs which gives the best connectivity for a given level of sparsity.

OBJECTIVE

Main contributions of our work are:

- RBGP Framework for generating Sparse Neural networks.
- RBGP4 Structured Sparse Pattern based on the above RBGP Framework.
- Utility of this work on various Computer Vision tasks and its results compared to benchmarks.

RBGP Framework

The core idea in RBGP (Ramanujan Bipartite Graph Product) framework is to express G as a bipartite graph product of Ramanujan bipartite graphs i.e ($G = G_1 \otimes_b \dots \otimes_b G_K$), where K is the number of base graphs. Advantages include:

- Structured Sparsity
- Memory Efficiency
- Good Connectivity

RBGP4 Sparsity Pattern

RBGP4 sparsity pattern constructed using four base Ramanujan bipartite graphs ($G = G_o \otimes_p Gr \otimes_p Gi \otimes_p G_b$), with graphs G_o and G_i being sparse, and G_r and G_b being complete bipartite graphs.

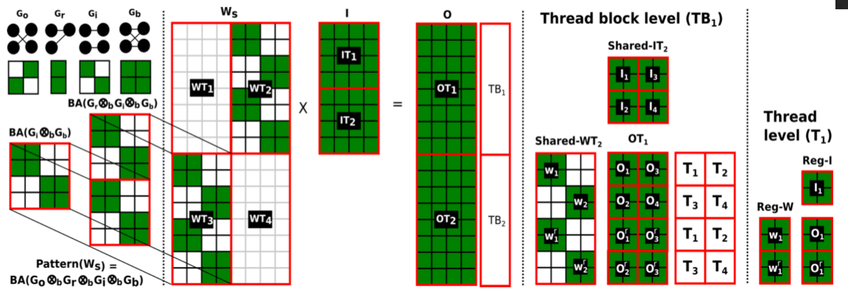
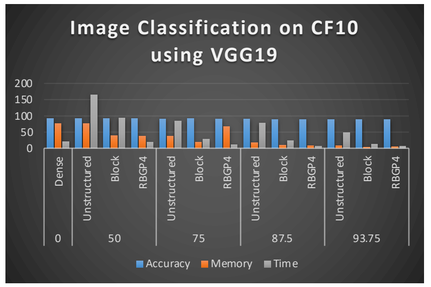
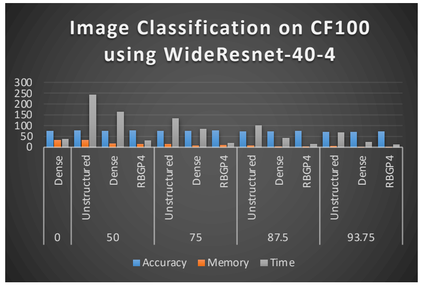


Figure 1: Tiled matrix multiplication of RBGP4 sparse matrix W_s with a dense matrix $I (O = W_s \times I)$

RESULTS

We have run this structure sparse network over various computer vision tasks like Image Classification on CIFAR Datasets.
We can clearly see a 5-9x reduction in time along with almost 2x reduction in memory usage while maintaining almost similar accuracy as a dense model.





FUTURE WORK

There are many possibilities to explore as this opens up the use of graph theory and also sparsity patterns.
We will also extend this work to larger datasets like Imagenet Classification and various ML tasks like Semantic Segmentation and TTS.

Authors: Dharma Teja Vooturi, Furqan Ahmed Shaik, Dr. Girish Varma, Dr. Kishore Kothapalli

Research Center Name: CSTAR