Systems for Machine Learning (CMSC828G)

Optimizing DL Kernels

Abhinav Bhatele, Daniel Nichols

UNIVERSITY OF MARYLAND

Announcements

- Assignment 2 is due on March 14 (extended to March 17 if you need the extra time)

GPU kernels in deep learning

- Important to focus on single GPU and single node performance before looking at scaling/distributed-memory performance
- Requires ensuring that GPU kernels execute as fast as possible
- Two research directions:
  - Systems optimizations - kernel fusion, reducing data movement, etc.
  - ML optimizations - changing the algorithm (for e.g. optimizer used)

Sparsity in deep learning

- Models can be pruned (zeroing out weights close to zero)
  - Reduces parameter counts
  - Lottery ticket hypothesis: sub-networks ("winning tickets") when trained in isolation reach test accuracies comparable to the original network
- Graph neural networks
  - Graph structure (vertices and edges) is represented as an adjacency matrix
- Scientific computing

GPUs not well-suited for sparse computations

- Irregular memory access patterns - not good for coalescing and prefetching memory accesses
- Sparse operations can result in conditional logic - not good for warp utilization
- Load imbalance across threads

Common sparse kernels

- SpMM: Sparse matrix multiply
  - multiplies a sparse matrix and a dense matrix
  - A is sparse and typically stored in Compressed Sparse Row (CSR) format, B is dense
- SDDMM: Sampled dense-dense matrix multiply
  - Element-wise dot product of a dense (AB) and sparse matrix (C)

UNIVERSITY OF
MARYLAND