Systems for Machine Learning (CMSC828G)

Sparsity in Deep Learning

Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Announcements

- Project proposal due date extended to March 17
- Office hours this week to discuss projects

What is sparsity?

- Traditional definition: refers to a elements in your data/structure being zero or close to zero
- Sparse matrices: when a large number of elements in a matrix are zero
  - Can happen because of data sparsity: recommendation models, graphs
- Model sparsity: Parameters that might be redundant and can be removed without sacrificing accuracy
- Structured vs unstructured sparsity

Why do we need sparsity?

- Reducing memory footprint
- Fewer parameters can also mean less computation

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)

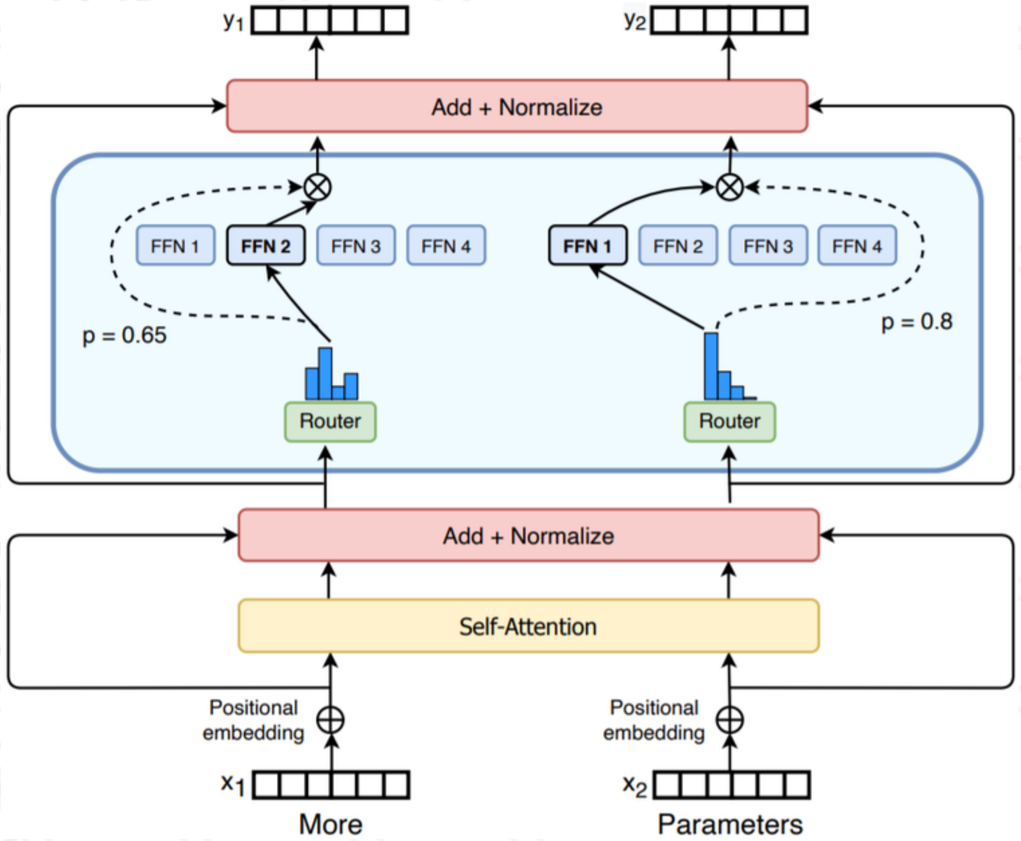
Pruning

- Remove weights close to zero
- Identify subnetworks (“winning tickets”) that are critical for model accuracy
- Not activate all weights all the time - Mixture-of-experts

Frankke et al. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. https://arxiv.org/abs/1803.03635

Mixture-of-experts

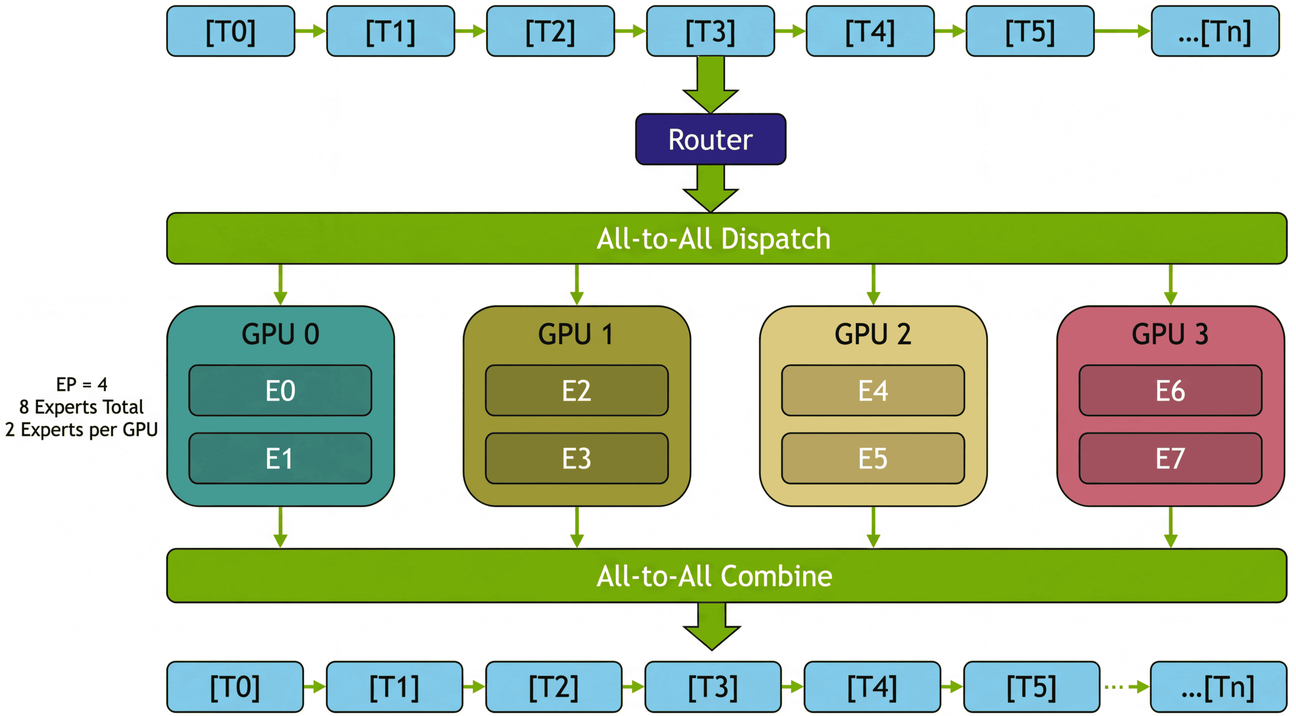
- Scale model capacity without scaling computation
- Only a subset of parameters (or "experts") are activated for each token
- Use "router" or gating network to decide which tokens should be sent to which experts



https://company.hpc-ai.com/blog/enhanced-moe-parallelism-open-source-moe-model-training-can-be-9-times-more-efficient

Expert parallelism

- Each GPU hosts a subset of experts
- All-to-all communication to send tokens to experts
- Load imbalance: if some experts are receiving more tokens



https://arxiv.org/abs/2603.07685

UNIVERSITY OF
MARYLAND