Tutorial. May 12, 2023 2-6 pm

Distributed Training of Deep Neural Networks

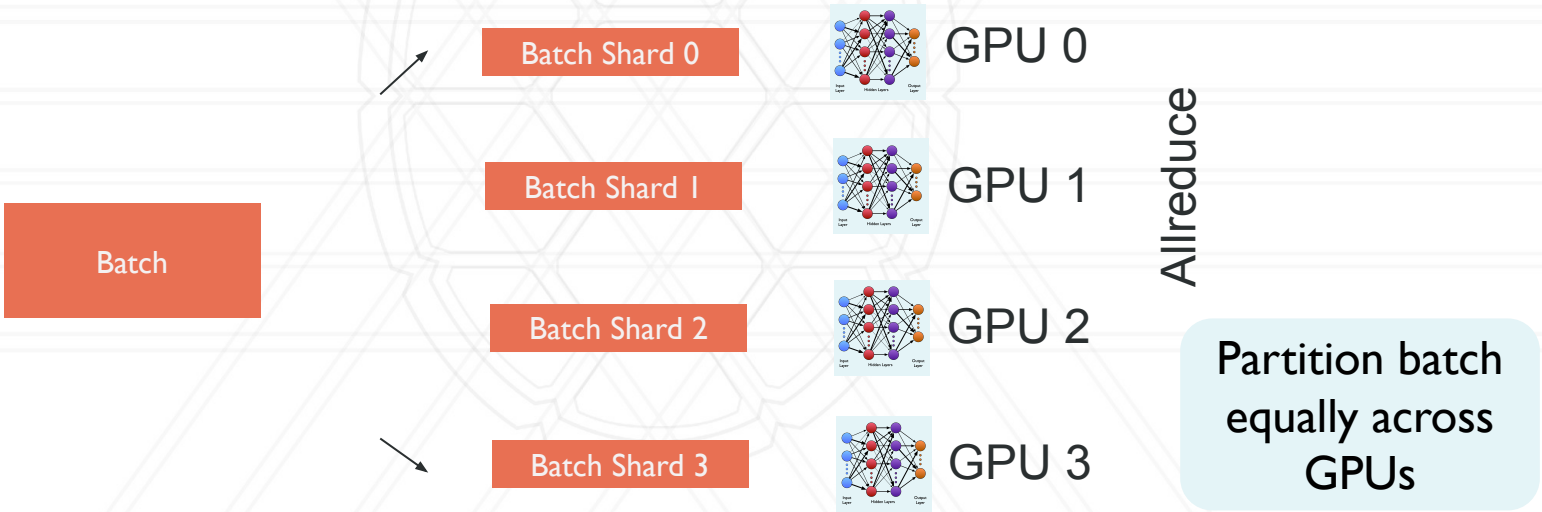
Abhinav Bhatele, Siddharth Singh, Daniel Nichols
Department of Computer Science

Data parallelism

- Work on different parts of the data in parallel on different GPUs
- Example: PyTorch's DDP, DeepSpeed's ZeRO

Work distribution in data parallelism

- Each worker has a full copy of the entire NN and processes different mini-batches
- All reduce operation to synchronize gradients



Using DDP

- Code location in the tutorial repo:
  session_2_data_parallelism/train_ddp.py

```
cd session_2_data_parallelism/
sbatch  --reservation=isc2024 run_ddp.sh
```

Using DeepSpeed

- Using DDP is limited to smaller model sizes
- ZeRO implements memory optimizations to fit larger models on a GPU
- Code location in the tutorial repo: session_2_data_parallelism/train_deepspeed.py

```
sbatch  --reservation=isc2024 run_deepspeed.sh
```

Abhinav Bhatele and Siddharth Singh
Department of Computer Science
bhatele@umd.edu, ssingh37@umd.edu