Critique on MegatronLM

Summary

The paper introduces methods to train large transformer models using intra-later model parallelism. The methods introduced does not require compiler or pipeline changes, rather it can be implemented through the insertion of a few communication operations in PyTorch. GEMMs are split across GPUs row-wise for the first matrix and column-wise for the second matrix. This can be extended to both MLP and Self-Attention Layers, and the output can be parallelized without any communication between devices.

The paper also shows that external parameters such as placement of layer normalization affects performance.

The paper also scales existing model sizes after playing around with the above hyperparams, and showcases SOTA results on datasets

Strengths

- Breaking down model architecture cuts down on memory required between each device, allowing for training of larger models
- Speeds up model training as GPUs can execute in parallel

Weaknesses

- Replication of activations and layer normalization values may lead to explosion in the memory space
- Communication is still required before each dropout layer

Suggested Improvements

- Addition of pipe lining to make sure time is better utilized while waiting for a mini-batch to complete execution