CSCI 1390, Spring 2025: Written HW 1

Read the following paper excerpts, and answer the questions below. Submit your answers as a single PDF on gradescope here.

Reading

- PipeDream – All Sections.
- Megatron-LM – All Sections.
- Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM – Sections 1, 3, Skim Evaluation.

Questions

The total length of your response should be about 500-600 words, with breakdowns specified below. Please adhere to the breakdowns; you will be penalized, for example, if your summarization is too long, and the other answers are too short. Additionally, we expect you to cite specific examples and evidence from the papers when answering the questions.

Summarization (200-300 words total)

1. What were the key challenges that the PipeDream paper solved to get pipeline parallelism working well in practice?
2. What are the advantages of tensor model parallelism over pipeline model parallelism; why was it chosen to train transformers?

Comprehension (200 words)

3. Why does the third paper explore combining parallelism strategies? What in this setting is different making it such that no one strategy alone is sufficient? Why is their method effective over any single parallelism strategy?

Synthesis (100-200 words)

4. Consider the workload of inference rather than training. Most transformer model weights will not fit on a single GPU. What parallelism strategy would you deploy and why?