Assignment One

Instructions

This assignment is self-contained. Please either typeset (or use any text-editor) to write your responses to your short answer question, attach files q2.py and q3.py alongside, and submit it to Canvas with your name and netid somewhere on your responses. For the coding assignments, please attach files q2.py and q3.py alongside the file containing the answers to the short answer questions.

1 Q1: Short Answer Questions – 10 pts

1. (1 points) Given a transformer with a model hidden dimension of 768 with 12 attention heads, what is the head hidden dimension?
2. (1 points) Given a transformer with a head hidden dimension of 28 and 29 attention heads, what is the model hidden dimension?
3. (2 points) During training, what data-type are the optimizer states stored in?
4. (3 points) Assume that we have a sequence length of s tokens. What is the time complexity of self-attention in the encoder and decoder (answer separately)?
5. (3 points) Given a standard MHA architecture with $L = 64$ layers, $d_{model} = 8192$ model hidden dimension, $h = 32$ heads per layer, $d_k = 256$ head dimension, and $n_{tokens} = 65536$, what is the total number of parameters and memory required to store the weights in bf16, assuming an MLP up-down projection factor of 4? Exclude biases, layernorm, and positional encoding parameters. Focus primarily on the token embeddings, attention, and MLP modules.

For the following two coding assignments, please refer to the repository's README.md for detailed instruction.

2 Q2: Parameter Server – 10 pts

Complete the code in q2.py to implement the parameter-server synchronization method.

- Server(5 pts)
- Worker(5 pts)

3 Q3: All-reduce – 30 pts

Complete the code in q3.py to implement the all-reduce synchronization method.

Point distribution:

- All-reduce(10pts)
- Reduce-scatter(10pts)
- All-gather(10pts)

Section Outline:

- Q1 Short Answer Questions – 10 pts
- Q2 Parameter Server – 10 pts
- Q3 All-reduce – 30 pts