COMP4901Y Homework2

Question 1. GPU Performance (20 points).

1.1 Note that the A100 GPU has:

- 1.41 GHz clock rate;
- There are 108 SM in one A100 GPU;
- One SM can process 512 TF32 FMA per clock.

What is the peak TF32 FLOPS? (5 points)

1.2 Suppose we have the following matrix multiplication computations in FP16. What is the corresponding arithmetic intensity? (15 points)

1. $Y = XW, X \in \mathbb{R}^{4096\times 4096}, W \in \mathbb{R}^{4096\times 4096}$
2. $Y = XW, X \in \mathbb{R}^{4096\times 32}, W \in \mathbb{R}^{32\times 4096}$
3. $Y = XW, X \in \mathbb{R}^{32\times 4096}, W \in \mathbb{R}^{4096\times 32}$

Submission. This part should be submitted with:

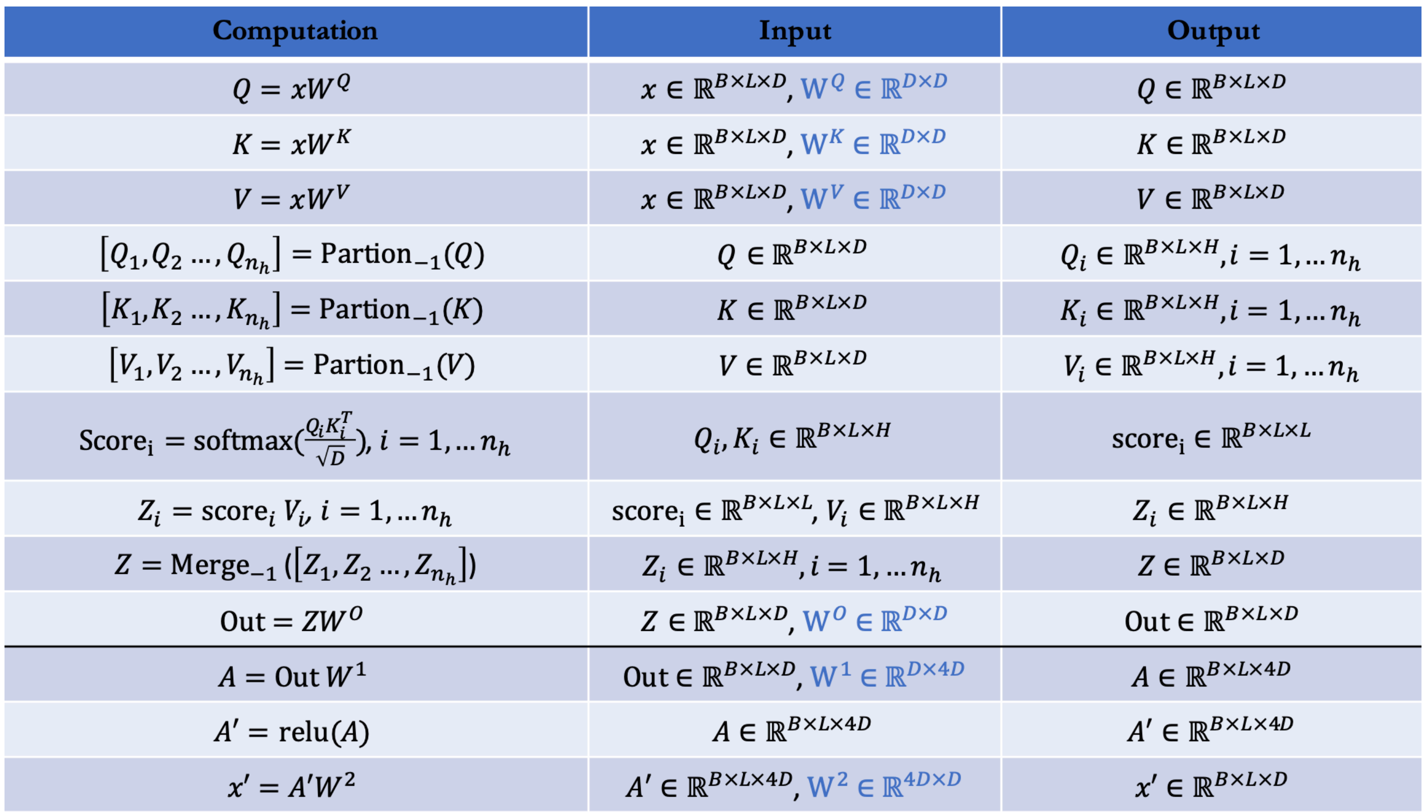
- A pdf file named **question1.pdf** to include the computation.

Question 2. Transformer and Parallel Training Estimation (40 points).

Given a model based on stacking transformer layers, where

- N_{layer} is the number of layers in the transformer layer;
- B is the training batch size;
- L is the training sequence length;
- D is the model dimension;
- n_H is the number of heads;
- H is the head dimension. Note that we have $D = n_H \times H$.

Each layer of computation is summarized below:



Computation	Input	Output
$Q = xW^Q$	$x \in \mathbb{R}^{B\times L\times D}$, $W^Q \in \mathbb{R}^{D\times D}$	$Q \in \mathbb{R}^{B\times L\times D}$
$K = xW^K$	$x \in \mathbb{R}^{B\times L\times D}$, $W^K \in \mathbb{R}^{D\times D}$	$K \in \mathbb{R}^{B\times L\times D}$
$V = xW^V$	$x \in \mathbb{R}^{B\times L\times D}$, $W^V \in \mathbb{R}^{D\times D}$	$V \in \mathbb{R}^{B\times L\times D}$
$[Q_1, Q_2 \dots, Q_{n_h}] = \text{Partion}_{-1}(Q)$	$Q \in \mathbb{R}^{B\times L\times D}$	$Q_i \in \mathbb{R}^{B\times L\times H}, i = 1, \dots n_h$
$[K_1, K_2 \dots, K_{n_h}] = \text{Partion}_{-1}(K)$	$K \in \mathbb{R}^{B\times L\times D}$	$K_i \in \mathbb{R}^{B\times L\times H}, i = 1, \dots n_h$
$[V_1, V_2 \dots, V_{n_h}] = \text{Partion}_{-1}(V)$	$V \in \mathbb{R}^{B\times L\times D}$	$V_i \in \mathbb{R}^{B\times L\times H}, i = 1, \dots n_h$
$\text{Score}_\text{i} = \text{softmax}(\frac{Q_i K_i^T}{\sqrt{D}}), i = 1, \dots n_h$	$Q_i, K_i \in \mathbb{R}^{B\times L\times H}$	$\text{score}_\text{i} \in \mathbb{R}^{B\times L\times L}$
$Z_i = \text{score}_i V_i, i = 1, \dots n_h$	$\text{score}_\text{i} \in \mathbb{R}^{B\times L\times L}$, $V_i \in \mathbb{R}^{B\times L\times H}$	$Z_i \in \mathbb{R}^{B\times L\times H}$
$Z = \text{Merge}_{-1}([Z_1, Z_2 \dots, Z_{n_h}])$	$Z_i \in \mathbb{R}^{B\times L\times H}, i = 1, \dots n_h$	$Z \in \mathbb{R}^{B\times L\times D}$
$\text{Out} = ZW^O$	$Z \in \mathbb{R}^{B\times L\times D}$, $W^O \in \mathbb{R}^{D\times D}$	$\text{Out} \in \mathbb{R}^{B\times L\times D}$
$A = \text{Out}\, W^1$	$\text{Out} \in \mathbb{R}^{B\times L\times D}$, $W^1 \in \mathbb{R}^{D\times 4D}$	$A \in \mathbb{R}^{B\times L\times 4D}$
$A' = \text{relu}(A)$	$A \in \mathbb{R}^{B\times L\times 4D}$	$A' \in \mathbb{R}^{B\times L\times 4D}$
$x' = A'W^2$	$A' \in \mathbb{R}^{B\times L\times 4D}$, $W^2 \in \mathbb{R}^{4D\times D}$	$x' \in \mathbb{R}^{B\times L\times D}$

Let us ignore all the other parts of the model (e.g., EmbedToken, position embedding, etc.) and make the following estimation:

Given a 7B model, where $N_{\text{layer}} = 32, B = 128, L = 4096, D = 4096, n_H = 32, H = 128$. Suppose all the computation is based on FP16. We have a cluster with 8 A100-80G GPUs.

2.1 Suppose we train the model by data parallelism (using the standard synchronous, lossless communication) implemented by **AllReduce.** For a training iteration, how many bytes in total should be aggregated through the **AllReduce** operations? You just need to specify the total bytes of the communication targets that are passed to the **AllReduce** API calls as input on each GPU. (10 points)

2.2 Suppose we train the model by pipeline parallelism implemented by Gpipe, where each stage handles 4 transformer layers. For a training iteration, how many bytes in total should be communicated between nearby stages i and $i+1$? Specify the peer-to-peer communication direction in your answer. (10 points)

2.3 Suppose we train the model by tensor model parallelism (where the tensor parallel degree is $D_{tp} = 8$). For a training iteration, how many bytes in total should be aggregated through the **AllReduce** operations? You just need to specify the total bytes of the communication targets that are passed to the **AllReduce** API calls as input on each GPU. (10 points)

2.4 Suppose we are training the model by fully sharded data parallelism. How many bytes in total should be communicated through the **AllGather** and **ReduceScatter** operations respectively? You just need to specify the total bytes of the communication targets that are passed to the **AllGather** or **ReduceScatter** API calls as input on each GPU. (10 points)

Submission. This part should be submitted with:

- A pdf file named **question2.pdf** to include the computation.

Question 3. Parallel Training Practice (40 points).

We will use the TACC platform to perform this part, and a tutorial will be hosted in the Lab on March 25, 2024. (Corresponding sample code will be released before the lab session.)

3.1 Data Parallelism Training in PyTorch. Run the sample code for DDP by testing the parallel degrees of $1, 2, 3, 4$ with the corresponding number of GPUs. Report the training duration for each of the settings. (20 points)

3.2 Fully Sharded Data Parallelism Training in PyTorch. Run the sample code for FSDP by testing the parallel degrees of $1, 2, 3, 4$ with the corresponding number of GPUs. Report the training duration for each of the settings. (20 points)

Submission. This part should be submitted with:

- A pdf file named **question3.pdf** to include the result table.

GPUs	1	2	3	4
DDP Training duration (seconds)				
FSDP Training duration (seconds)				

Submission Checklist.

You should submit a zip file including the following components:

- **question1.pdf**
- **question2.pdf**
- **question3.pdf**