COMP4901Y Homework3

Question 1. Transformer Generative Inference Computation (40 points).

Given a model based on stacking transformer layers, where

- L is the input sequence length;
- D is the model dimension;
- n_H is the number of heads;
- H is the head dimension. Note that we have $D = n_H \times H$.

Suppose $L = 4096, D = 4096, n_H = 32, H = 128$.

1.1 Compute the arithmetic intensity of the matrix multiplication during the prefill phase by filling the following form (steps 1,2,3,7,8,9,11,13, for step 7 only consider $Q_i K_i^T$, ignore the rest part.) [20 points]

No.	Computation	Input	Output	Arithmetic Intensity
1	$Q = xW^Q$	$x \in \mathbb{R}^{L\times D}$, $W^Q \in \mathbb{R}^{D\times D}$	$Q \in \mathbb{R}^{L\times D}$	
2	$K = xW^K$	$x \in \mathbb{R}^{L\times D}$, $W^K \in \mathbb{R}^{D\times D}$	$K \in \mathbb{R}^{L\times D}$	
3	$V = xW^V$	$x \in \mathbb{R}^{L\times D}$, $W^V \in \mathbb{R}^{D\times D}$	$V \in \mathbb{R}^{L\times D}$	
4	$[Q_1, Q_2 \dots, Q_{n_h}] = \text{Partition}_{-1}(Q)$	$Q \in \mathbb{R}^{L\times D}$	$Q_i \in \mathbb{R}^{L\times H}, i = 1, \dots n_h$	-
5	$[K_1, K_2 \dots, K_{n_h}] = \text{Partition}_{-1}(K)$	$K \in \mathbb{R}^{L\times D}$	$K_i \in \mathbb{R}^{L\times H}, i = 1, \dots n_h$	-
6	$[V_1, V_2 \dots, V_{n_h}] = \text{Partition}_{-1}(V)$	$V \in \mathbb{R}^{L\times D}$	$V_i \in \mathbb{R}^{L\times H}, i = 1, \dots n_h$	-
7	$\text{Score}_\text{i} = \text{softmax}(\frac{Q_i K_i^T}{\sqrt{D}}), i = 1, \dots n_h$	$Q_i, K_i \in \mathbb{R}^{L\times H}$	$\text{score}_\text{i} \in \mathbb{R}^{L\times L}$	
8	$Z_i = \text{score}_i V_i, i = 1, \dots n_h$	$\text{score}_\text{i} \in \mathbb{R}^{L\times L}$, $V_i \in \mathbb{R}^{L\times H}$	$Z_i \in \mathbb{R}^{L\times H}$	
9	$Z = \text{Merge}_{-1}([Z_1, Z_2 \dots, Z_{n_h}])$	$Z_i \in \mathbb{R}^{L\times H}, i = 1, \dots n_h$	$Z \in \mathbb{R}^{L\times D}$	-
10	$\text{Out} = ZW^O$	$Z \in \mathbb{R}^{L\times D}$, $W^O \in \mathbb{R}^{D\times D}$	$\text{Out} \in \mathbb{R}^{L\times D}$	
11	$A = \text{Out}\, W^1$	$\text{Out} \in \mathbb{R}^{L\times D}$, $W^1 \in \mathbb{R}^{D\times 4D}$	$A \in \mathbb{R}^{L\times 4D}$	
12	$A' = \text{relu}(A)$	$A \in \mathbb{R}^{L\times 4D}$	$A' \in \mathbb{R}^{L\times 4D}$	-
13	$x' = A'W^2$	$A' \in \mathbb{R}^{L\times 4D}$, $W^2 \in \mathbb{R}^{4D\times D}$	$x' \in \mathbb{R}^{L\times D}$	

1.2 Compute the arithmetic intensity of the matrix multiplication during the decode phase (generate the first token) by filling the following form (steps 1,2,4,9,10,12,13,15, for step 9 only consider $Q_i K_i^T$, ignore the rest part.) [20 points.]

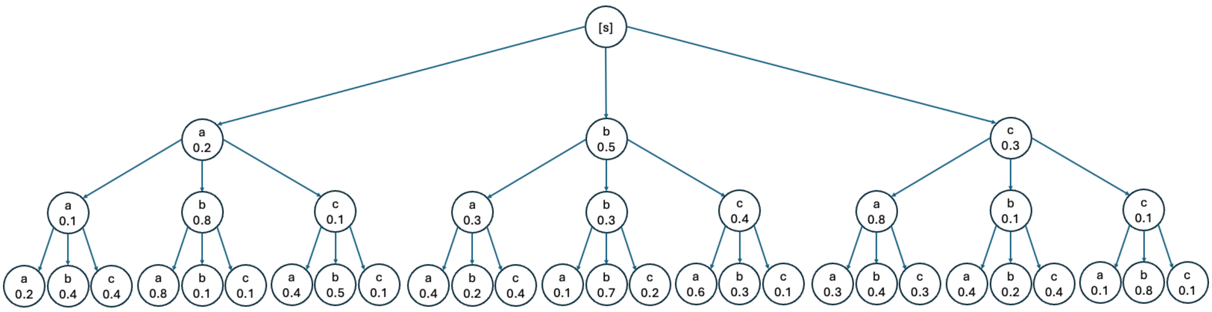
No	Computation	Input	Output	Arithmetic Intensity
1	$Q = Q_d = tW^Q$	$t \in \mathbb{R}^{1\times D}$, $W^Q \in \mathbb{R}^{D\times D}$	$Q, Q_d \in \mathbb{R}^{1\times D}$	
2	$K_d = tW^K$	$t \in \mathbb{R}^{1\times D}$, $W^K \in \mathbb{R}^{D\times D}$	$K_d \in \mathbb{R}^{1\times D}$	
3	$K = \text{concat}(K_{\text{cache}}, K_d)$	$K_{\text{cache}} \in \mathbb{R}^{L\times D}, K_d \in \mathbb{R}^{1\times D}$	$K \in \mathbb{R}^{(L+1)\times D}$	-
4	$V_d = tW^V$	$t \in \mathbb{R}^{1\times D}$, $W^V \in \mathbb{R}^{D\times D}$	$V_d \in \mathbb{R}^{1\times D}$	
5	$V = \text{concat}(V_{\text{cache}}, V_d)$	$V_{\text{cache}} \in \mathbb{R}^{L\times D}, V_d \in \mathbb{R}^{1\times D}$	$V \in \mathbb{R}^{(L+1)\times D}$	-
6	$[Q_1, Q_2 \dots, Q_{n_h}] = \text{Partition}_{-1}(Q)$	$Q \in \mathbb{R}^{1\times D}$	$Q_i \in \mathbb{R}^{1\times H}, i = 1, \dots n_h$	-
7	$[K_1, K_2 \dots, K_{n_h}] = \text{Partition}_{-1}(K)$	$K \in \mathbb{R}^{(L+1)\times D}$	$K_i \in \mathbb{R}^{(L+1)\times H}, i = 1, \dots n_h$	-
8	$[V_1, V_2 \dots, V_{n_h}] = \text{Partition}_{-1}(V)$	$V \in \mathbb{R}^{(L+1)\times D}$	$V_i \in \mathbb{R}^{(L+1)\times H}, i = 1, \dots n_h$	-
9	$\text{Score}_i = \text{softmax}(\frac{Q_i K_i^T}{\sqrt{D}}), i = 1, \dots n_h$	$Q_i \in \mathbb{R}^{1\times H}, K_i \in \mathbb{R}^{(L+1)\times H}$	$\text{score}_i \in \mathbb{R}^{1\times(L+1)}$	
10	$Z_i = \text{score}_i\, V_i, i = 1, \dots n_h$	$\text{score}_i \in \mathbb{R}^{1\times(L+1)}, V_i \in \mathbb{R}^{(L+1)\times H}$	$Z_i \in \mathbb{R}^{1\times H}$	
11	$Z = \text{Merge}_{-1}([Z_1, Z_2 \dots, Z_{n_h}])$	$Z_i \in \mathbb{R}^{1\times H}, i = 1, \dots n_h$	$Z \in \mathbb{R}^{1\times D}$	-
12	$\text{Out} = ZW^O$	$Z \in \mathbb{R}^{1\times D}$, $W^O \in \mathbb{R}^{D\times D}$	$\text{Out} \in \mathbb{R}^{1\times D}$	
13	$A = \text{Out}\, W^1$	$\text{Out} \in \mathbb{R}^{1\times D}$, $W^1 \in \mathbb{R}^{D\times 4D}$	$A \in \mathbb{R}^{1\times 4D}$	
14	$A' = \text{relu}(A)$	$A \in \mathbb{R}^{1\times 4D}$	$A' \in \mathbb{R}^{1\times 4D}$	-
15	$t' = A'W^2$	$A' \in \mathbb{R}^{1\times 4D}$, $W^2 \in \mathbb{R}^{4D\times D}$	$t' \in \mathbb{R}^{1\times D}$	

Submission. This part should be submitted with:

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

Question 2. Beam Search (20 points).

Given a vocabulary set of {a, b, c}, we have the exhaustive search of the generation track below:



Suppose we use the beam search with the number of beams as 2. Please mark the beam search track in the graph and write down the finally generated 2 candidate sequences.

Submission. This part should be submitted with:

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

Question 3. Hugging Face Inference Test (40 points).

We will use the hugging face API to learn how to do generative inference computation easily. Use the implementation provided in question3.py, and change the arguments when you start the Python script to test different cases:

3.1 Change the output token number by changing --output_len [10 points];

3.2 Play with beam search by changing --beam_width [10 points];

3.3 Play with the top-k sampling by changing --top_k [1o points];

3.4 Play with the top-k sampling by changing --top_p [1o points].

Submission. This part should be submitted with:

- A PDF file named **question3.pdf** that includes your launch command and the generated output for each of the 4 sub-tasks.
- Note that you can change the prompt input by changing —prompt_seq if you like, but do not use inappropriate content, including pornography, violence, extreme political and religious opinions, etc.