Flash Attention

Question 1

0 / 1 pts

Tri Dao mentions in the talk that softmax computation can be broken down into smaller independent pieces as follows.

How can the parameters like alpha and beta be obtained?

Tiling

Decomposing large softmax into smaller ones by scaling.

$$\text{softmax}([A_1, A_2]) = [\alpha\ \text{softmax}(A_1),\ \ \beta\ \text{softmax}(A_2)]\,.$$

$$\text{softmax}([A_1, A_2]) \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \alpha\ \text{softmax}(A_1)\ V_1 + \beta\ \text{softmax}(A_2)\ V_2\,.$$

swer

- Maintain additional statistics such as m and l

ered

- They come from pre-processing where matrices are broken down

- alpha and beta only rely on local information, so they can be computed completely independently

- Perform final normalization in CPU

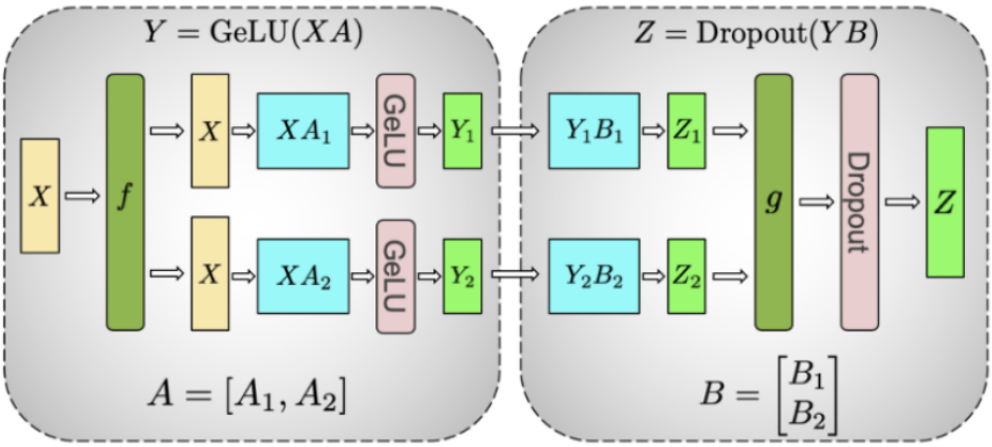
Megatron LM

Question 1

0 / 1 pts

The following diagram shows how Megatron-LM splits its computation for the feed-forward layer.

Why not compute Y as [X1 X2] * [A1 A2]^T ?



- Because X is very small in practice
- Because GeLU is non-linear
- Because we need to gather them for Z anyway
- Because CUDA is faster with duplicate input

DeepSpeed ZeRo

Question 1

1 / 1 pts

The diagram below indicates how ZeRO would partition parameters across multiple GPUs. Which of the following is INCORRECTLY describing the approach?

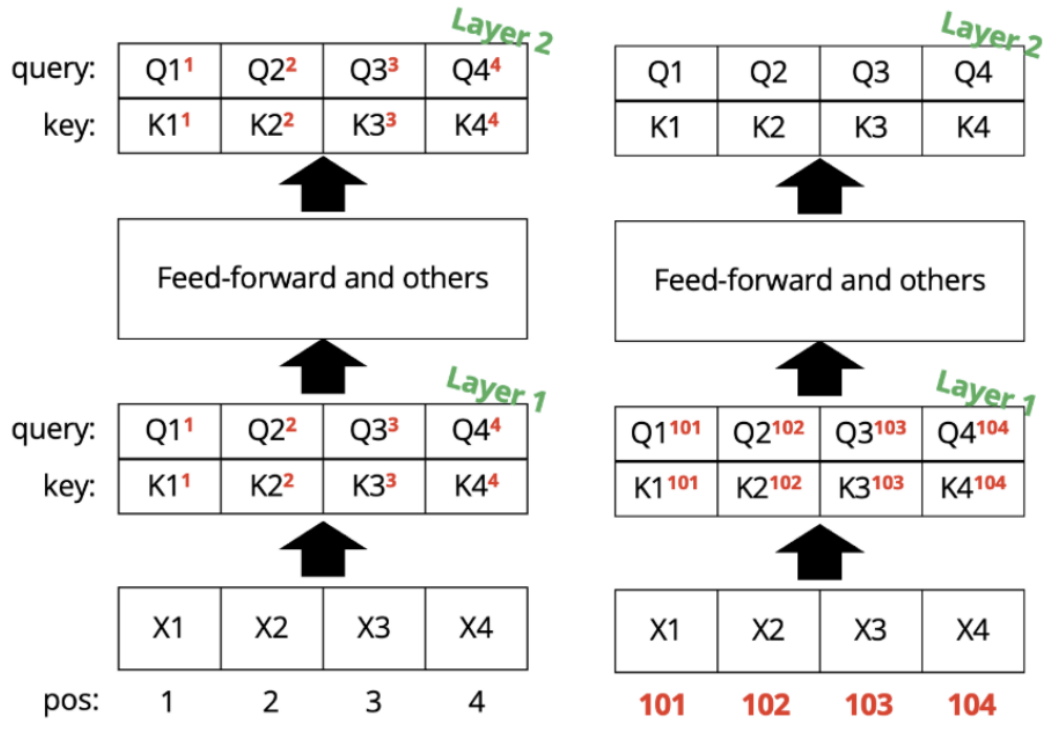
- If there are N parameters, each GPU holds N/4 parameters
- ct! Model weights are not broadcasted
- Green boxes are for Adam
- Gradient values are updated after completing forward pass

Block Attention

Question 1

1 / 1 pts

The following diagram depicts position re-encoding of KV states. When we move KV states from position 1-4 to 101-104, which of the following should be performed?



- Rotate query states
- Rotate value states
- Re-compute part of feed-forward
- Re-compute part of attention scores
- Rotate key states

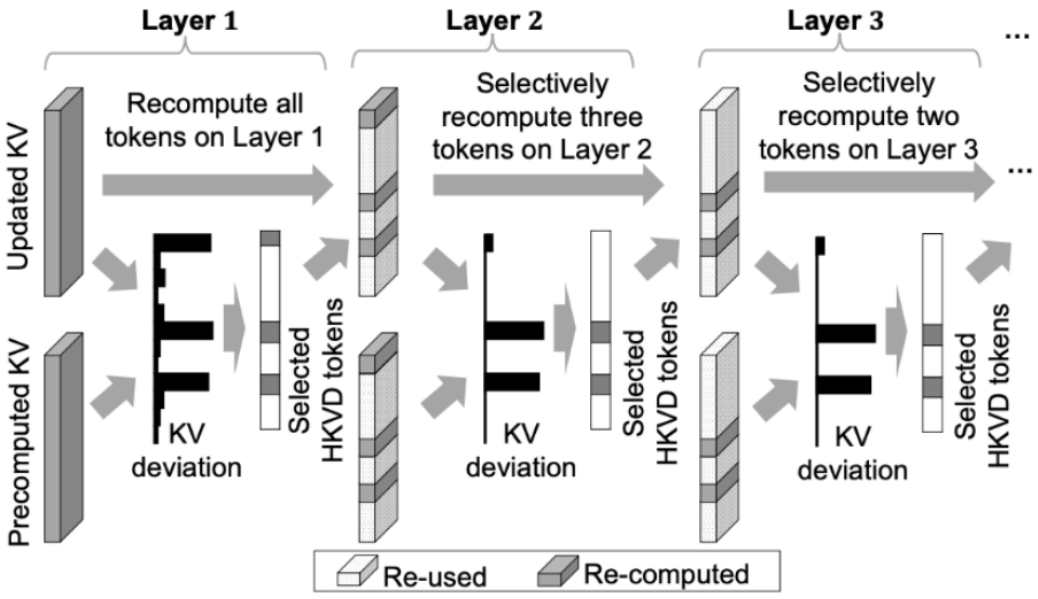
1 backlin

Cache Blend

Question 1

1 / 1 pts

CacheBlend proposes ***KV deviation*** to recompute cross attention selectively. Which of the following layers (among Layer 1, 2, and 3) are most crucial in computing KV deviation?



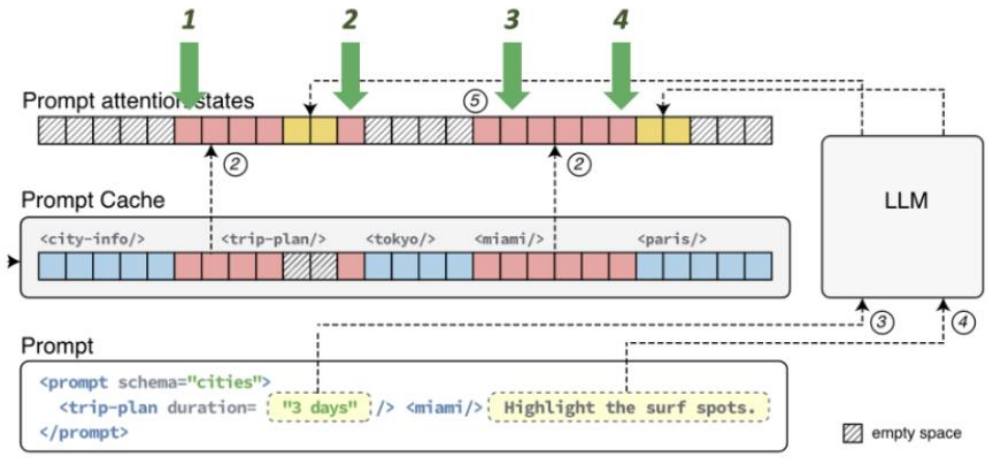
- Layer 3
- Layer 2
- Layer 1

Prompt Cache

Question 1

1 / 1 pts

Prompt Cache facilitates the reuse of KV cache. While red tokens are supposed to be simple copies from the cache, we discussed that some may need to be recomputed for accurate attention. Which one is that?



- 1
- 3
- 2
- 4

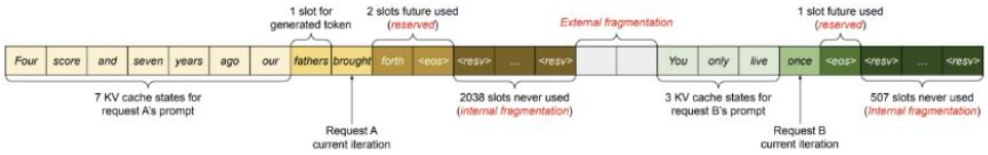
t!

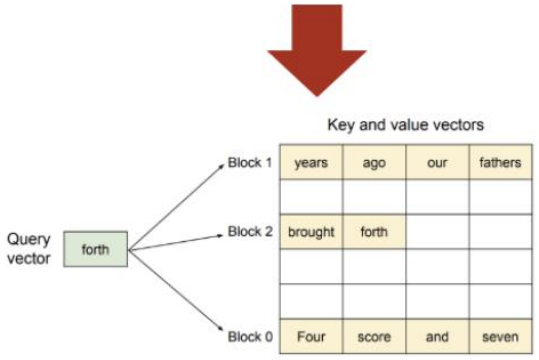
vLLM Paged Attention

Question 1

1 / 1 pts

Yellow boxes indicate the slots used for Request 1. How many slots are saved by using PagedAttention?





- 12
- 2048
- 2
- 2036

ct!

Kishu

Question 1

0 / 1 pts

Suppose I have the following Python code. Which of the following is a proper Co-variable?

```
a = 1
b = 2
c = 3
mylist1 = [a, b]
mylist2 = [b, c]
```

swered

- (mylist1, mylist2)
- (mylist1, a, b)
- (mylist2, b, c)
- No correct answers on this list

Answer

- (a, b, c mylist1, mylist2)

Orca

Question 1

1 / 1 pts

For efficient iterations, Orca aims to merge as many operations as possible. Which of the following operations can be merged between two prompts (P1 and P2)?

ct!

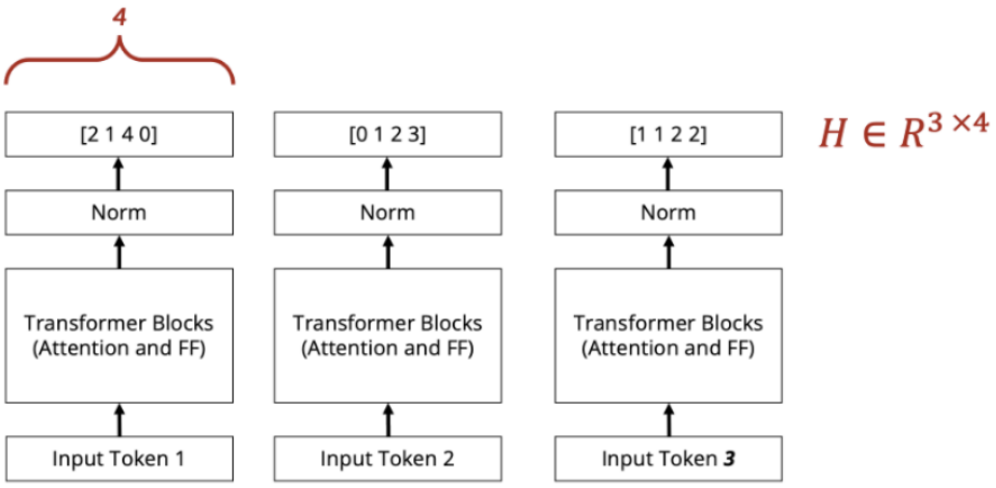
- (●) P1's feed-forward (for token 10) and P2's feed-forward (for token 11)
- ( ) P1's decoding (for token 10) and P2's decoding (for token 11)
- ( ) P1's prefill and P2's decoding
- ( ) P1's prefill (length 10) and P2's prefill (length 20)

LLM Embeddings

Question 1

1 / 1 pts

We aim to obtain an embedding through mean pooling. Which of the following will be the output?

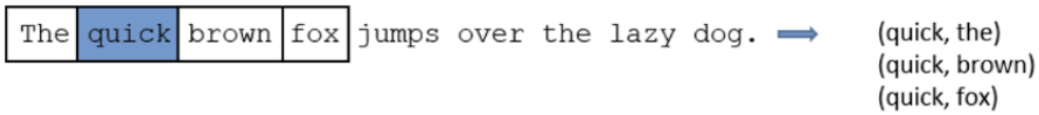


- [2 1 4 0]
- [2 1 4 0 0 1 2 3 1 1 2 2]
- [1 1 2.67 1.67]
- [1 1 2 2]

Question 1

1 / 1 pts

According to this training data, which of the following words is LEAST LIKELY to appear around the word "quick".



- fox
- brown
- jumps
- The