Homework 11

Anshul Sawant

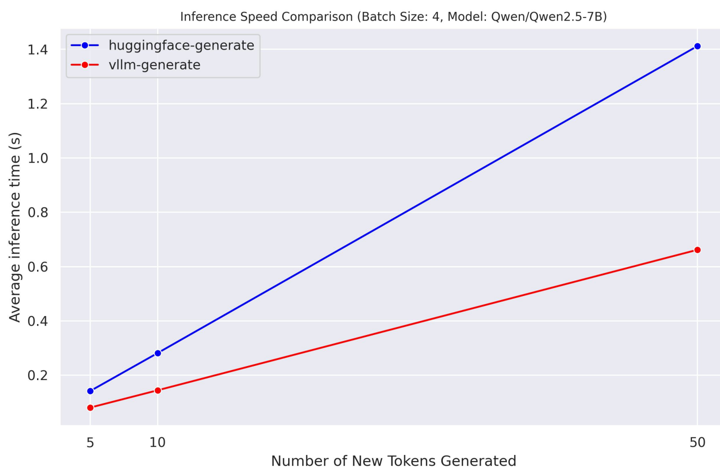
April 14, 2025

Contents

1 Q 1.1

2 A (coding)

3 B



The inference is faster with vLLM primarily because of PagedAttention and continuous batchingt.

4 Q 1.2

Flash attention makes sense as it improves the speed of forward pass.

Data parallelism Overall the technique is not applicable. Only the model replication aspect applies.

Gradient checkpointing No. No gradients are involved in the forward pass.

Deepspeed zero No. Gradients and optimizer states are not required for inference.

5 Q 1.3

5.1 A

Allocating large chunk is not effective because:

1. Wasted memory as most inference requests will require less K-V cache memory leading to lower throughput.

2. Techniques that generate multiple output sequences (such as beam search) can potentially share parts (corresponding to the input sequence) of K-V cache. However, this cannot be done because of the seperate pre-allocation for each sequence.

5.2 B

The dynamic size of the KV cache during inference contrasts with static model components, demanding a different memory management approach. PagedAttention provides this by adapting operating system paging techniques. It uses fixed-size memory blocks and indirection tables to manage the KV cache efficiently, mirroring how an OS creates a contiguous virtual memory view over fragmented physical pages, thus enabling flexible allocation.

5.3 C

One can possibly use the input/output of the unquantized model as the calibration dataset.

5.4 D

Reduced range can lead to overflow and underflow leading to infinity/NaNs which can propagate throughout the computation.