Online Result Summary

Model: yolov5tensorrt

GPU(s): NVIDIA GeForce RTX 3070 Laptop GPU

Total Available GPU Memory: 7.8 GB

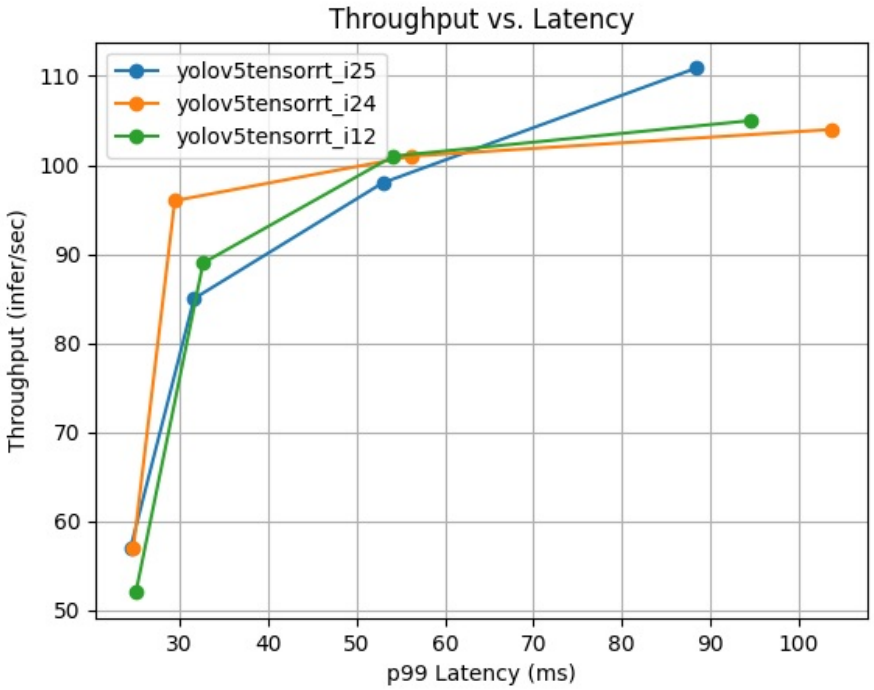
Client Request Batch Size: 1

Constraint targets: None

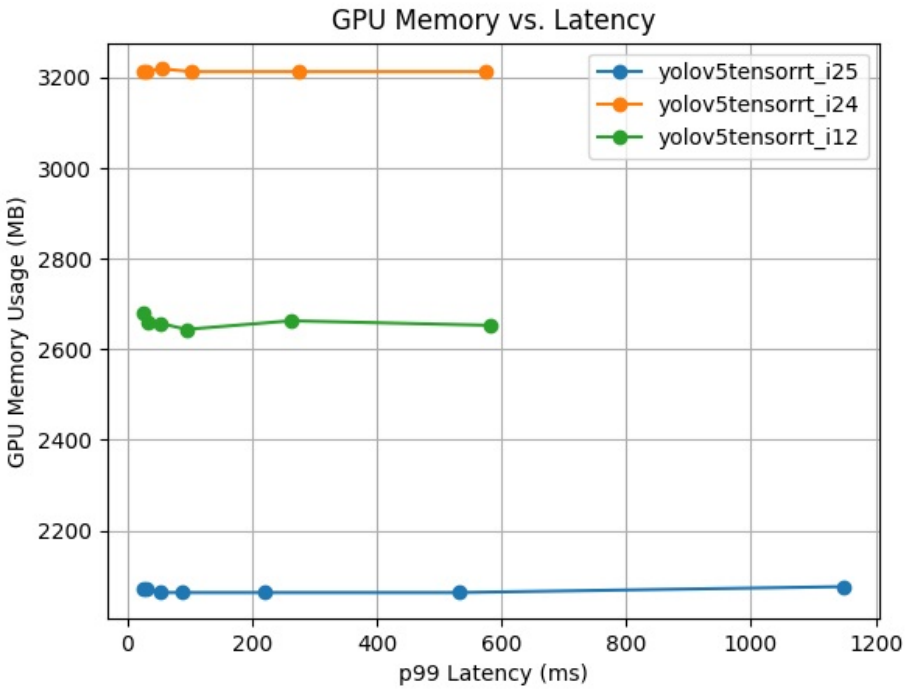
In 193 measurement(s), 1/GPU model instance(s) with preferred batch size of [8] on platform tensorrt_plan delivers maximum throughput under the given constraints on GPU(s) NVIDIA GeForce RTX 3070 Laptop GPU.

Curves corresponding to the 3 best model configuration(s) out of a total of 30 are shown in the plots.

The maximum GPU memory consumption for each of the above points is shown in the second plot. The GPUs NVIDIA GeForce RTX 3070 Laptop GPU have a total available memory of 7.8 GB respectively.



Throughput vs. Latency curves for 3 best configurations.



GPU Memory vs. Latency curves for 3 best configurations.

The following table summarizes each configuration at the measurement that optimizes the desired metrics under the given constraints.

Model Config Name	Preferred Batch Size	Instance Count	p99 Latency (ms)	Throughput (infer/sec)	Max CPU Memory Usage (MB)	Max GPU Memory Usage (MB)	Average GPU Utilization (%)
yolov5tensorrt_i25	[8]	1/GPU	24.474	57.0	0	2071.0	15.9
yolov5tensorrt_i24	[4]	5/GPU	24.796	57.0	0	3215.0	13.0
yolov5tensorrt_i12	[1]	3/GPU	25.071	52.0	0	2681.0	16.3