Online Result Summary

Model: unet34

GPU(s): 1 x NVIDIA GeForce RTX 3050 Ti Laptop GPU

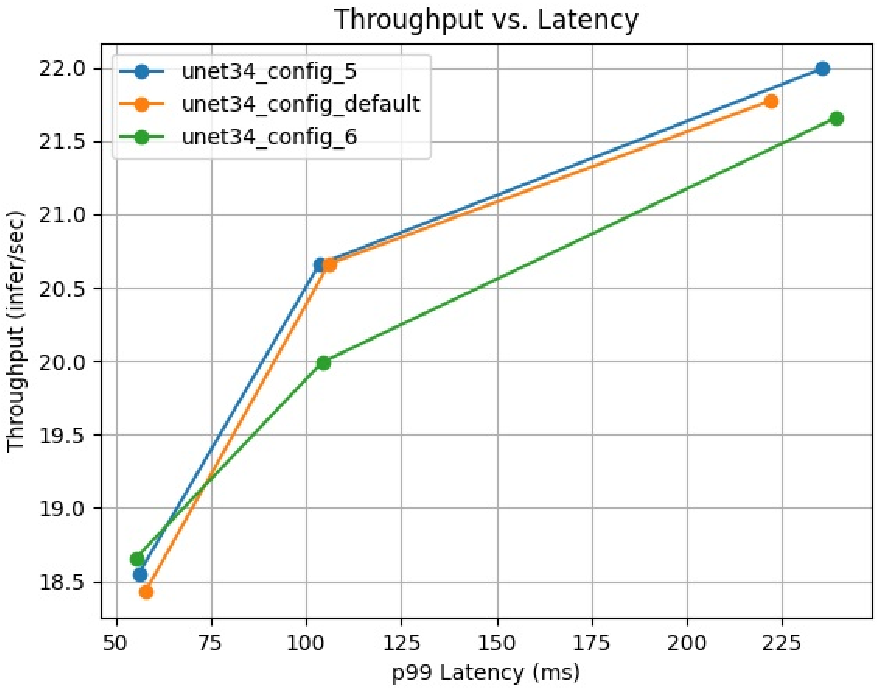
Total Available GPU Memory: 3.8 GB

Constraint targets: None

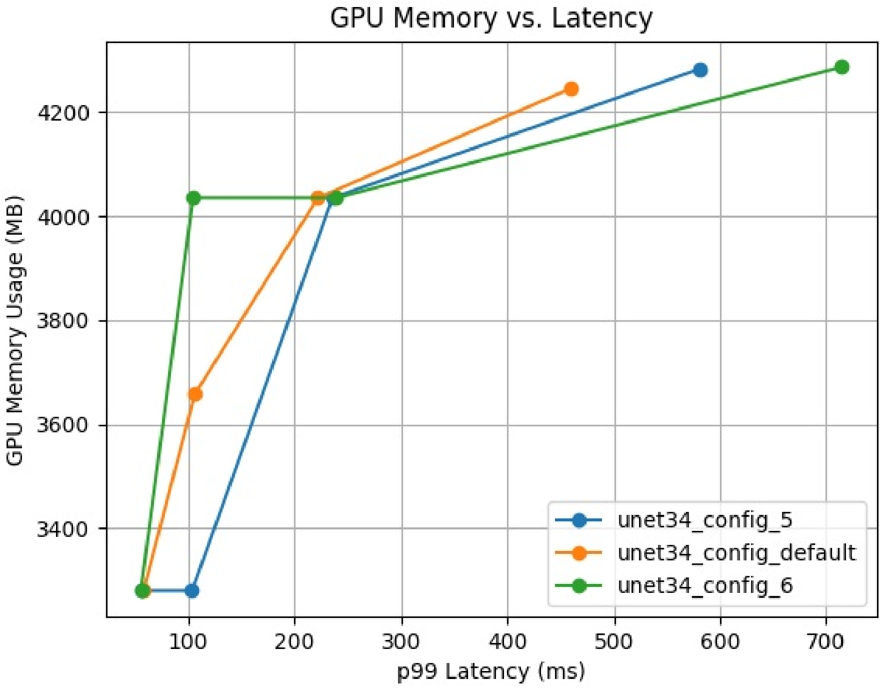
In 42 measurements across 16 configurations, **unet34_config_5** is **1%** better than the default configuration at meeting the objectives, under the given constraints, on GPU(s) 1 x NVIDIA GeForce RTX 3050 Ti Laptop GPU.

- **unet34_config_5**: 2 GPU instances with a max batch size of 4 on platform pytorch_libtorch

Curves corresponding to the 3 best model configuration(s) out of a total of 16 are shown in the plots.



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	**Max Batch Size**	**Dynamic Batching**	**Total Instance Count**	**p99 Latency (ms)**	**Throughput (infer/sec)**	**Max GPU Memory Usage (MB)**	**Average GPU Utilization (%)**
unet34_config_5	4	Enabled	2:GPU	235.588	21.9929	4034	100.0
unet34_config_default	6	Enabled	2:GPU	221.899	21.7725	4034	90.0
unet34_config_6	8	Enabled	2:GPU	239.376	21.6599	4034	89.2