Online Result Summary

Model: densenet_onnx

GPU(s): 1 x NVIDIA GeForce RTX 3060 Laptop GPU

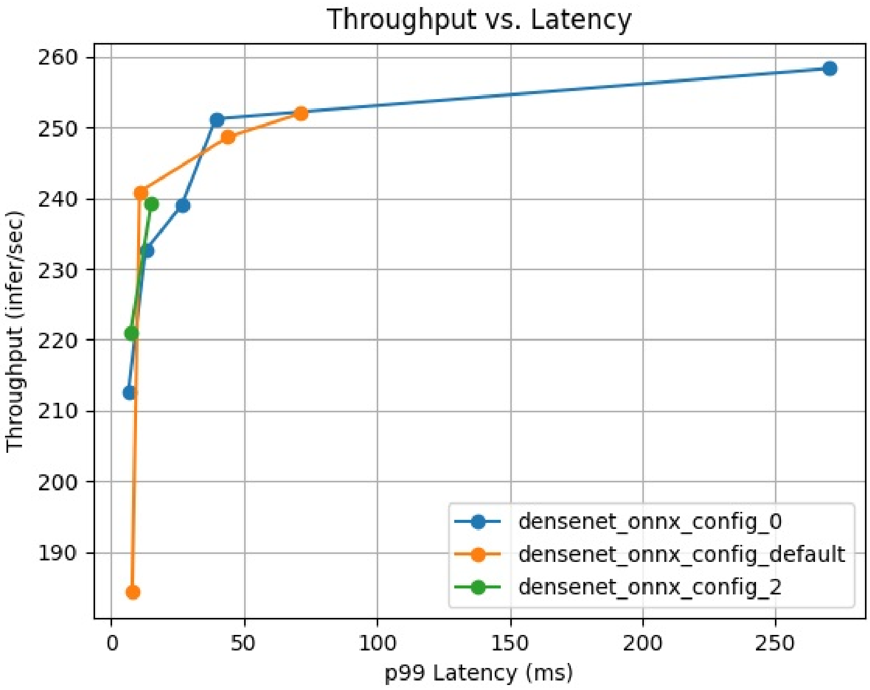
Total Available GPU Memory: 6.0 GB

Constraint targets: None

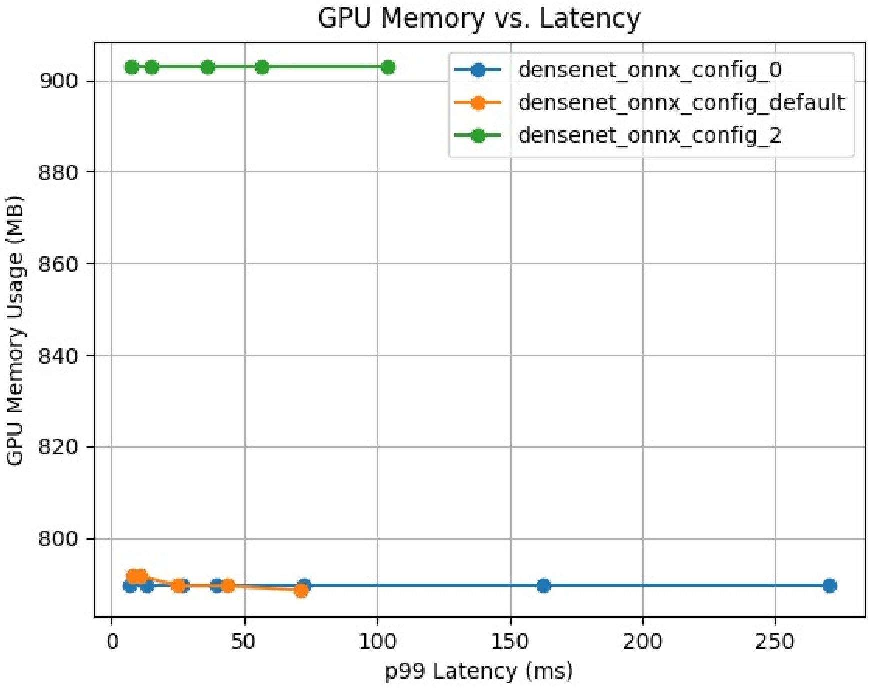
In 31 measurements across 6 configurations, **densenet_onnx_config_0** is **3%** better than the default configuration at maximizing throughput, under the given constraints, on GPU(s) 1 x NVIDIA GeForce RTX 3060 Laptop GPU.

- **densenet_onnx_config_0**: 1 GPU instance with a max batch size of 0 on platform onnxruntime

Curves corresponding to the 3 best model configuration(s) out of a total of 6 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 (%)
densenet_onnx_config_0	0	Enabled	1:GPU	270.471	258.316	789	75.3
densenet_onnx_config_default	0	Enabled	1:GPU	71.122	251.949	788	37.0
densenet_onnx_config_2	0	Enabled	3:GPU	15.079	239.332	902	57.0