AI Deployment Report

Best Optimization (Default Config)

Congratulations! You are able to boost deployment performance up to **11.3X** on your model with the most performant

Ranking	Optimization Set	Performance (sample/sec)
1	Intel Neural Compressor Post-Training Static Quantization (FX)	756
2	Intel Neural Compressor Post-Training Static Quantization (FX) + Channels Last + TorchDynamo JIT Script	753
3	Intel Neural Compressor Post-Training Static Quantization (FX) + TorchDynamo JIT Script	751
52	Default	67

*(1) All optimization sets are measured with default configuration (single instance on single socket), among which the top 3 performant ones are displayed. (2) This report evaluates performance only (accuracy under development).

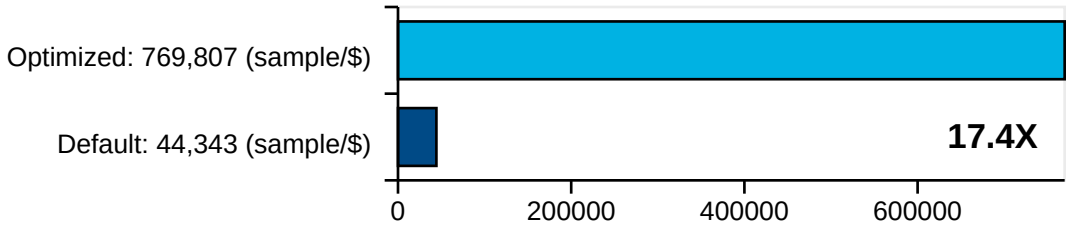
Best Optimization (Sweeping Configs)

For the most performant optimization set, you can further boost your deployment performance to up to **17.4X** if using the most performant deployment configuration according to our sweeping result.

Category	Num Intances	Num Cores Per Instance	BS	Performance (sample/sec)
Throughput	4	8	32	1163
Throughput based on P50-Latency	1	32	64	1368
Throughput based on P90-Latency	1	32	64	1367
Throughput based on P99-Latency	1	32	64	1367

*Measured on the most performant optimization set (Ranking 1 in above table) by sweeping configurations among batch size, number of instances, and number of cores per instance.

Cost Saving



*Sample/$ is calculated based on AWS c6i.32xlarge instance and on-demand price.