efficientnet_b0 Inference Latency vs Batch Size