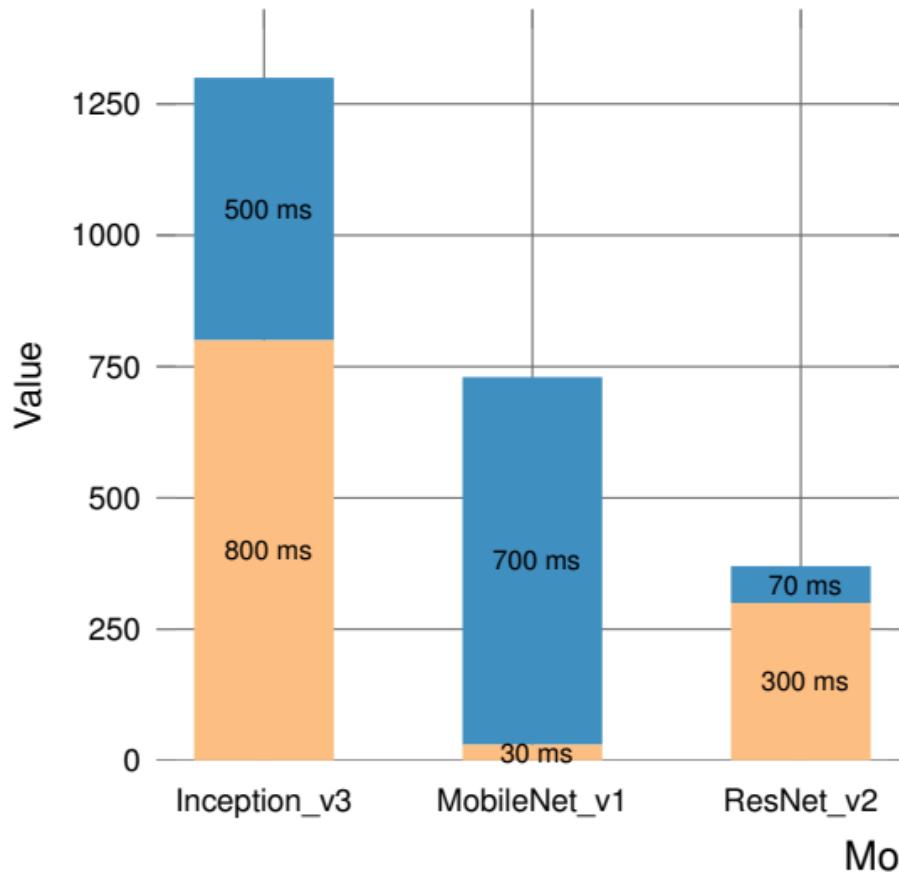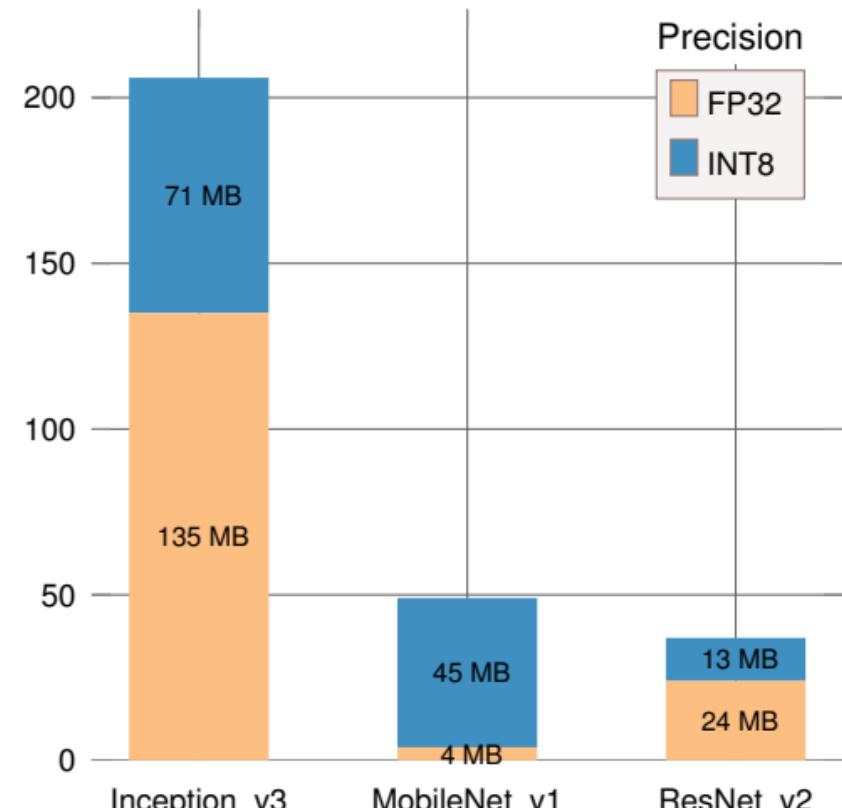Impact of Quantization on Inference Time and Model Size