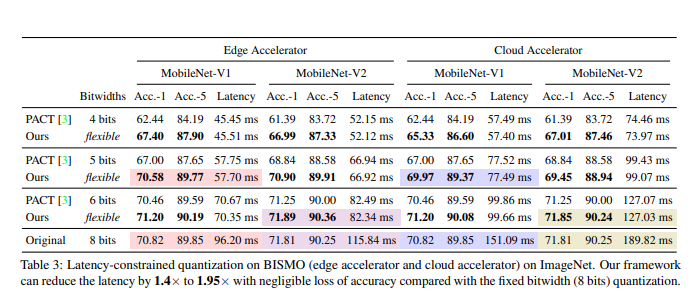
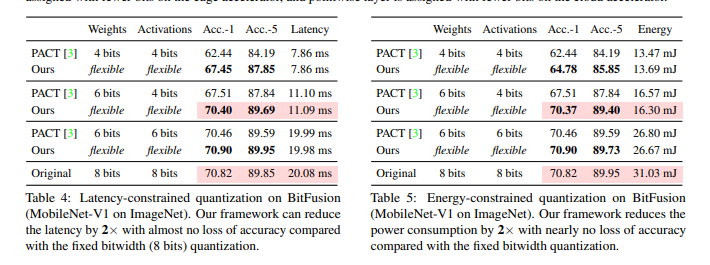
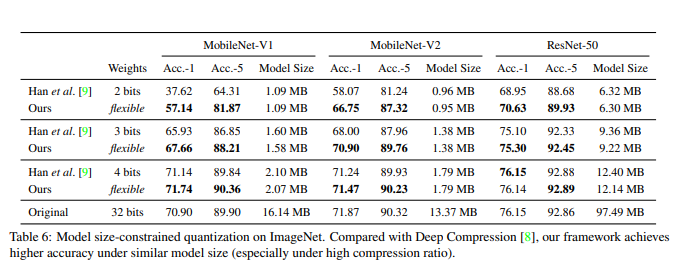
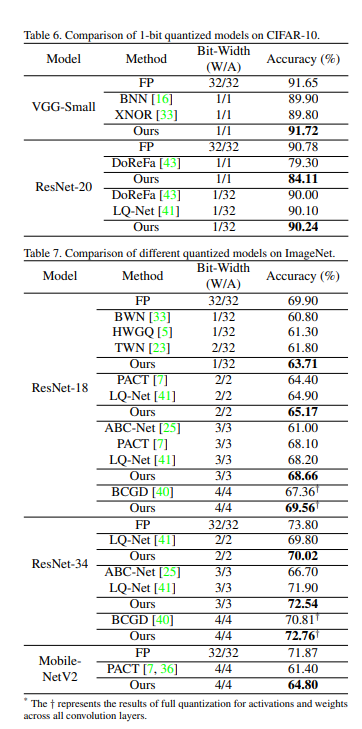
HAQ: Hardware-Aware Automated Quantization with Mixed Precision



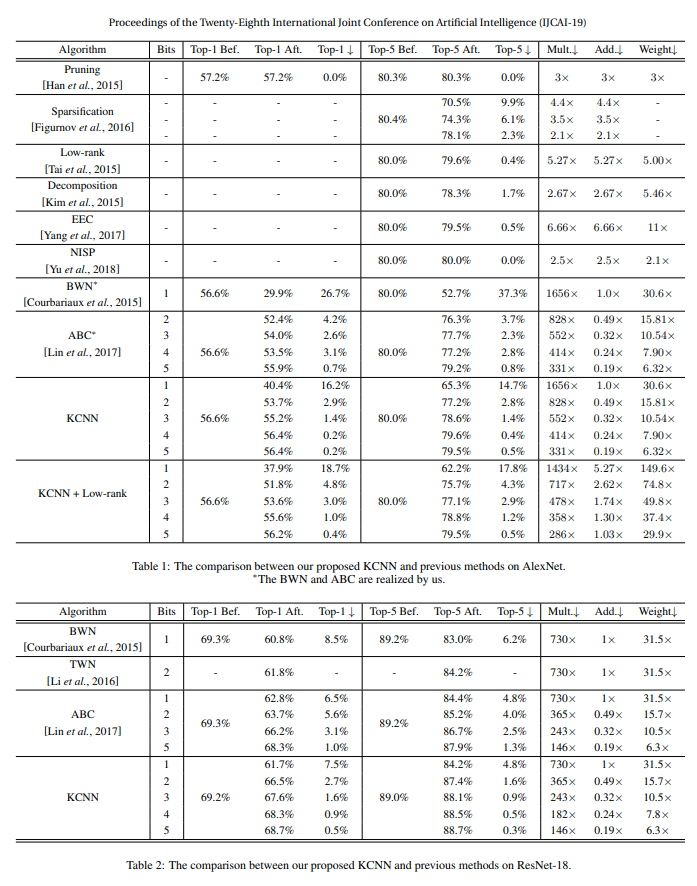




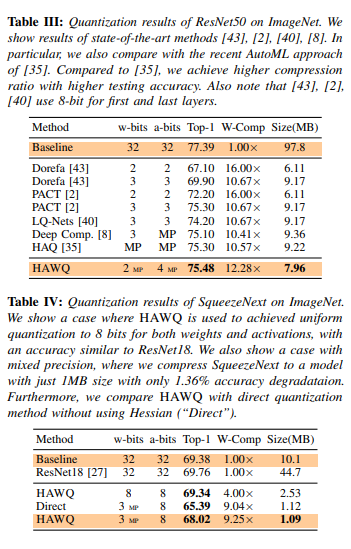
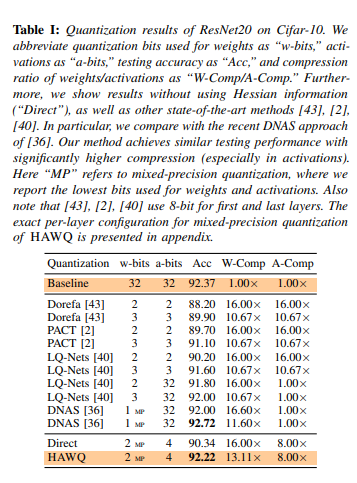
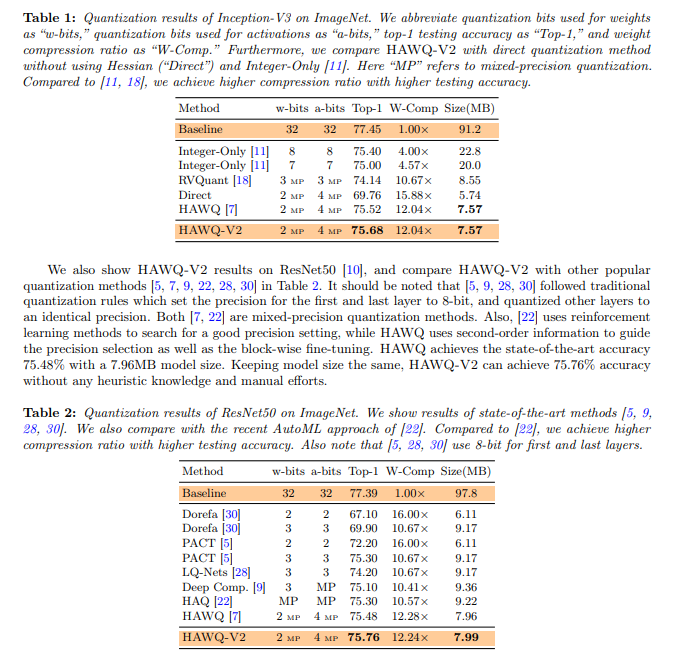
Differentiable Soft Quantization: Bridging Full-Precision and Low-Bit Neural Networks

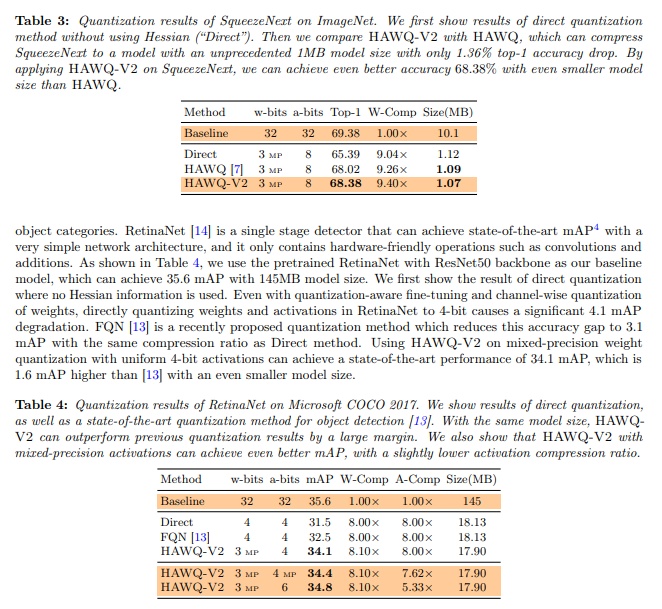


KCNN: Kernel-wise Quantization to Remarkably Decrease Multiplications in Convolutional Neural Network

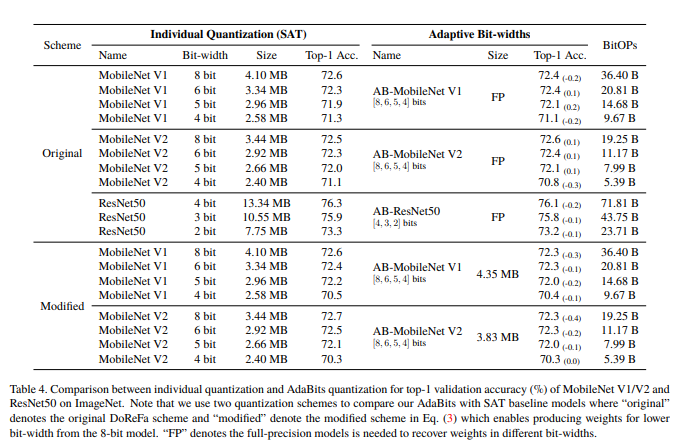


HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision

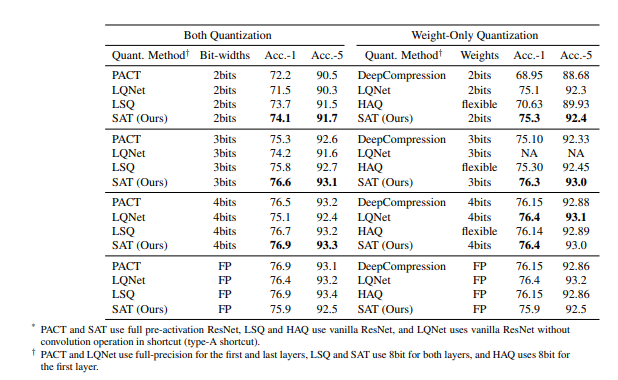
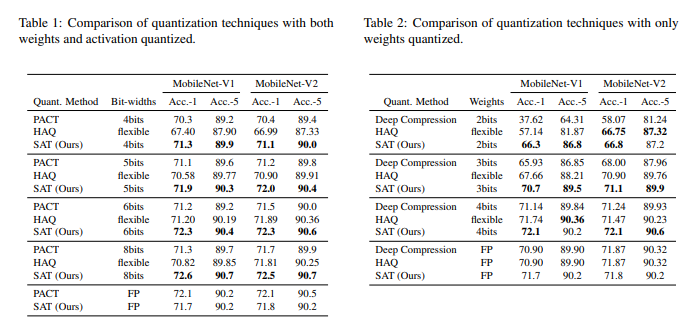
HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks



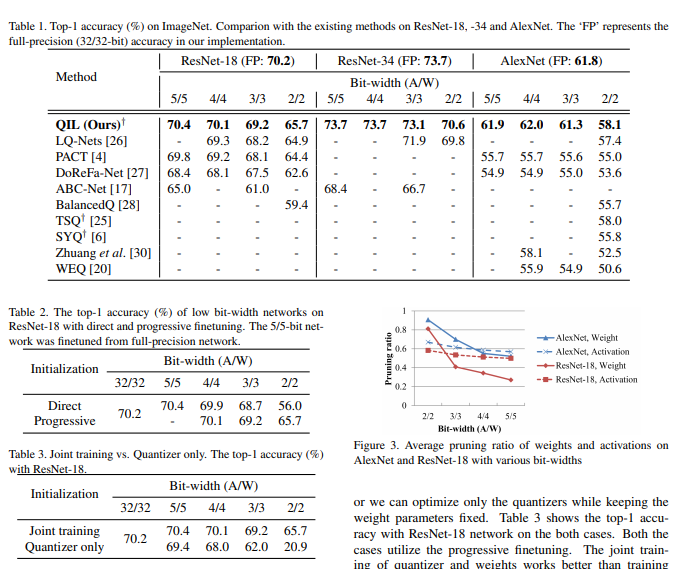
AdaBits: Neural Network Quantization with Adaptive Bit-Widths



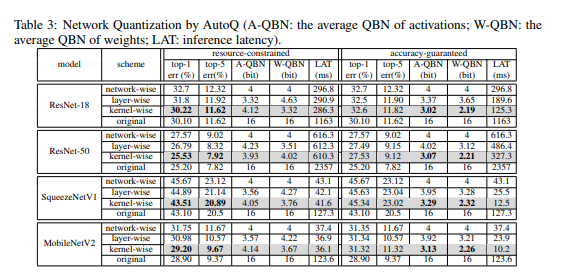
Towards Efficient Training for Neural Network Quantization



Learning to Quantize Deep Networks by Optimizing Quantization Intervals with Task Loss



AUTOQ: AUTOMATED KERNEL-WISE NEURAL NETWORK QUANTIZATION∗



MIXED PRECISION DNNS: ALL YOU NEED IS A GOOD PARAMETRIZATION

