Model quantization is a widely used technique to compress and accelerate deep neural network (DNN) inference. Emergent DNN hardware accelerators begin t