PACT: PARAMETERIZED CLIPPING ACTIVATION FOR QUANTIZED NEURAL NETWORKS

2018-11-02

Overview

- 2018
- 截断的ReLU函数，截断的大小可以训练
- 实验和对比很充分，但是感觉idea略显简单

PACT

$$y = PACT(x) = 0.5(|x| - |x - \alpha| + \alpha) = \begin{cases} 0, & x \in (-\infty, 0) \\ x, & x \in [0, \alpha) \\ \alpha, & x \in [\alpha, +\infty) \end{cases} \tag{1}$$

where α limits the range of activation to $[0, \alpha]$. The truncated activation output is then linearly quantized to k bits for the dot-product computations, where

$$y_q = round(y \cdot \frac{2^k - 1}{\alpha}) \cdot \frac{\alpha}{2^k - 1} \tag{2}$$

With this new activation function, α is a variable in the loss function, whose value can be optimized during training. For back-propagation, gradient $\frac{\partial y_q}{\partial \alpha}$ can be computed using the Straight-Through Estimator (STE) (Bengio et al. (2013)) to estimate $\frac{\partial y_q}{\partial y}$ as 1. Thus,

$$\frac{\partial y_q}{\partial \alpha} = \frac{\partial y_q}{\partial y}\frac{\partial y}{\partial \alpha} = \begin{cases} 0, & x \in (-\infty, \alpha) \\ 1, & x \in [\alpha, +\infty) \end{cases} \tag{3}$$

The larger the α, the more the parameterized clipping function resembles a ReLU Actfn. To avoid

PACT: PARAMETERIZED CLIPPING ACTIVATION FUNCTION

- 对alpha进行L2正则，为了让范围尽可能小一点，下图可以看到正则系数为0.01时各层的alpha的变化情况：最后普遍接近1和2

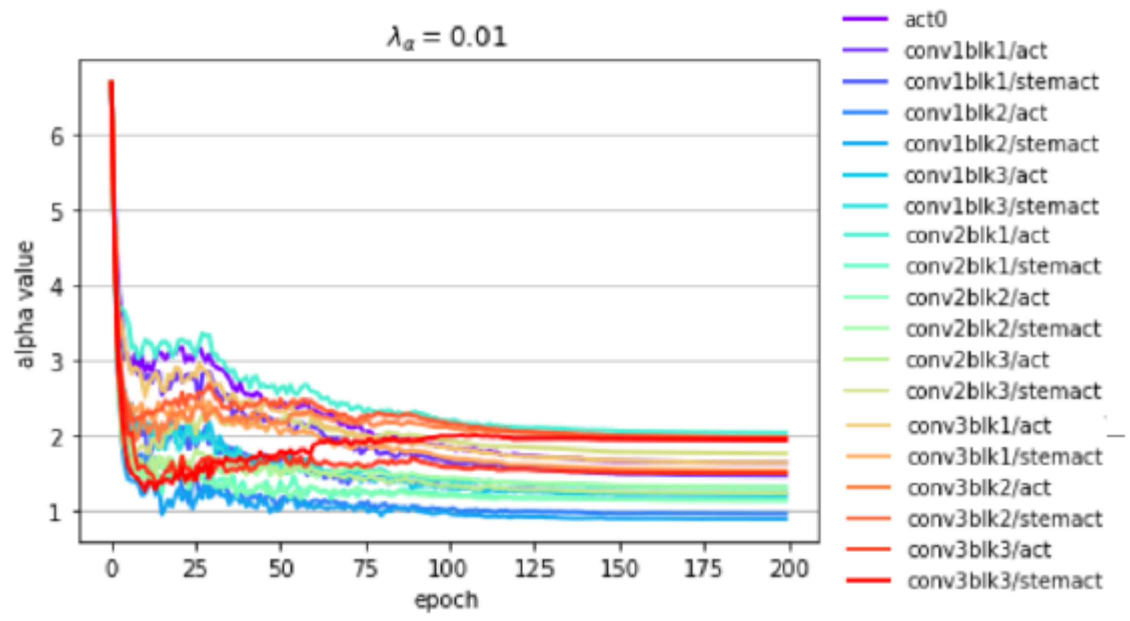
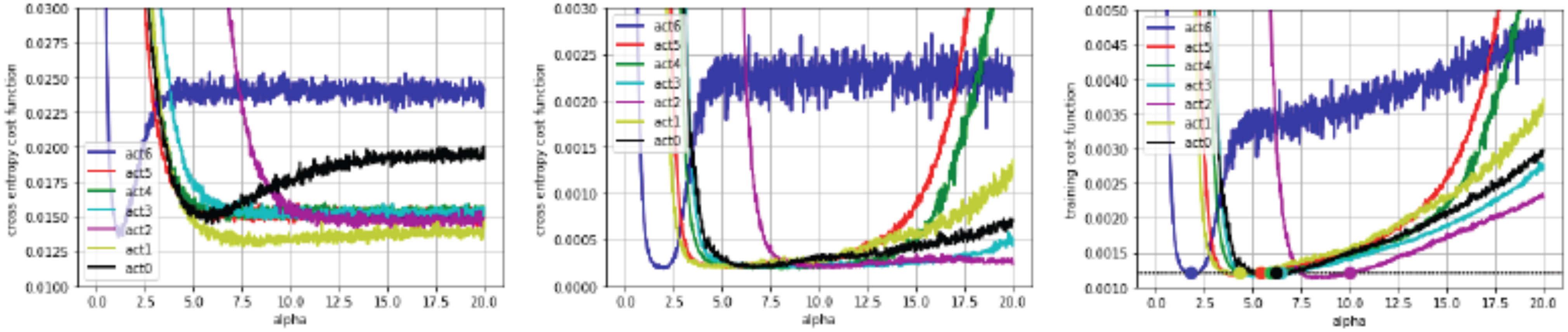


Figure 2: Evolution of α values during training using a ResNet20 model on the CIFAR10 dataset.

UNDERSTANDING HOW PARAMETERIZED CLIPPING WORKS

- 比较改变alpha时，量化和不量化时对应的loss，使用pretrain模型，对应实验值改变一个变量，第一个说明不做量化时，alpha的值也能大幅的减少loss的值，量化的结果也是，对于不同的层也有不同的最好的alpha值，说明了学习的重要性



(a) Cross-entropy in full-precision (b) Cross-entropy with quantization (c) Training loss

Figure 3: Cross-entropy vs α for SVHN image classification.

EXPLORATION OF HYPER-PARAMETERS

- 每层共用一个alpha即可，效果最好

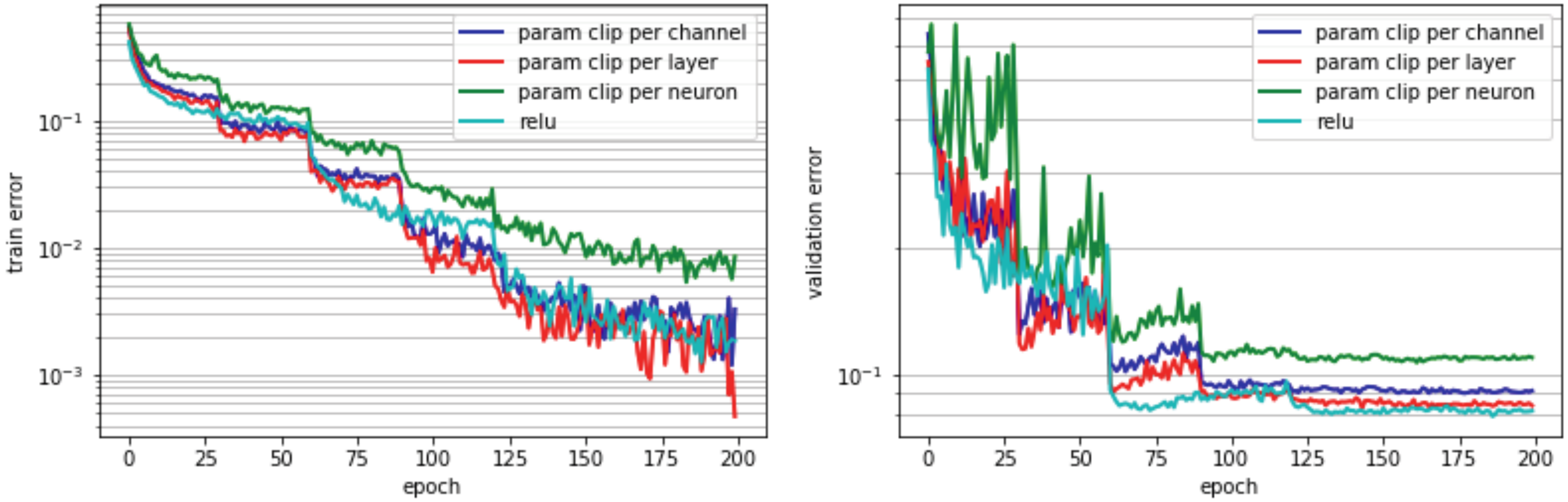


Figure 7: Training and validation error of CIFAR10-ResNet20 for PACT with different scope of α.

EXPLORATION OF HYPER-PARAMETERS

- 初始化alpha时，相对大一些比较好，因为太小会梯度消失
- L2系数和weight共用就可以，改变系数也不会太影响精度

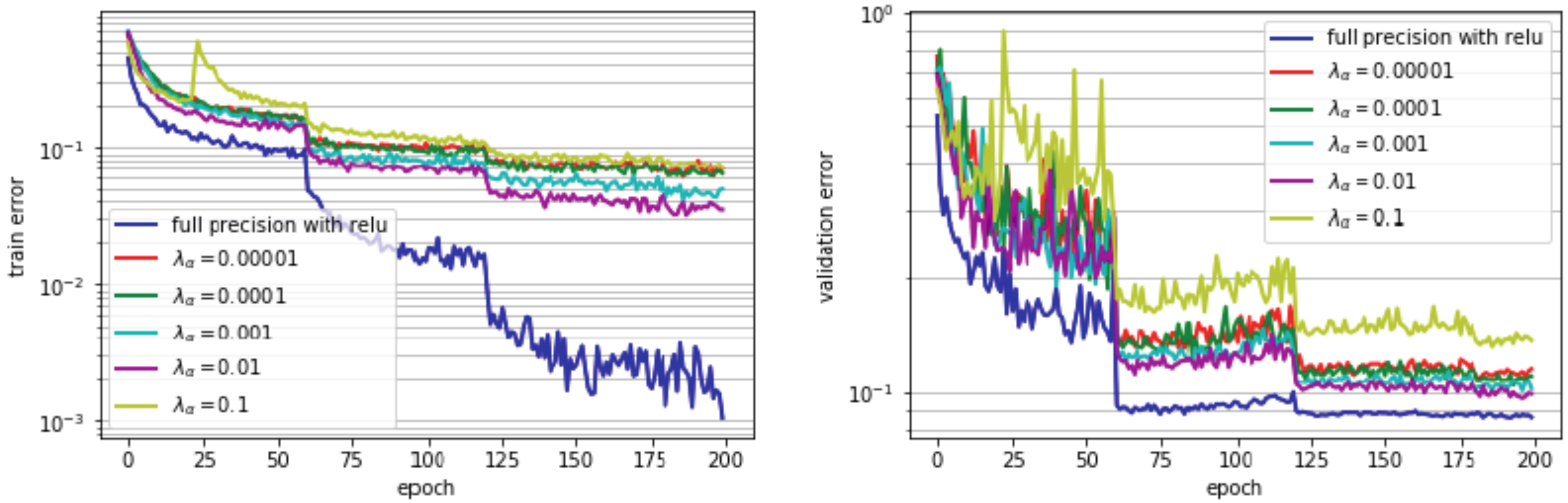


Figure 8: Training and validation error of quantized CIFAR10-ResNet20 for PACT with different regularization parameter λ_α.

EXPLORATION OF HYPER-PARAMETERS

- 不量化第一层和最后一层，但是实验发现第一层和最后一层量化8-bit时精度损失是最小的，所以其实可以用8-bit

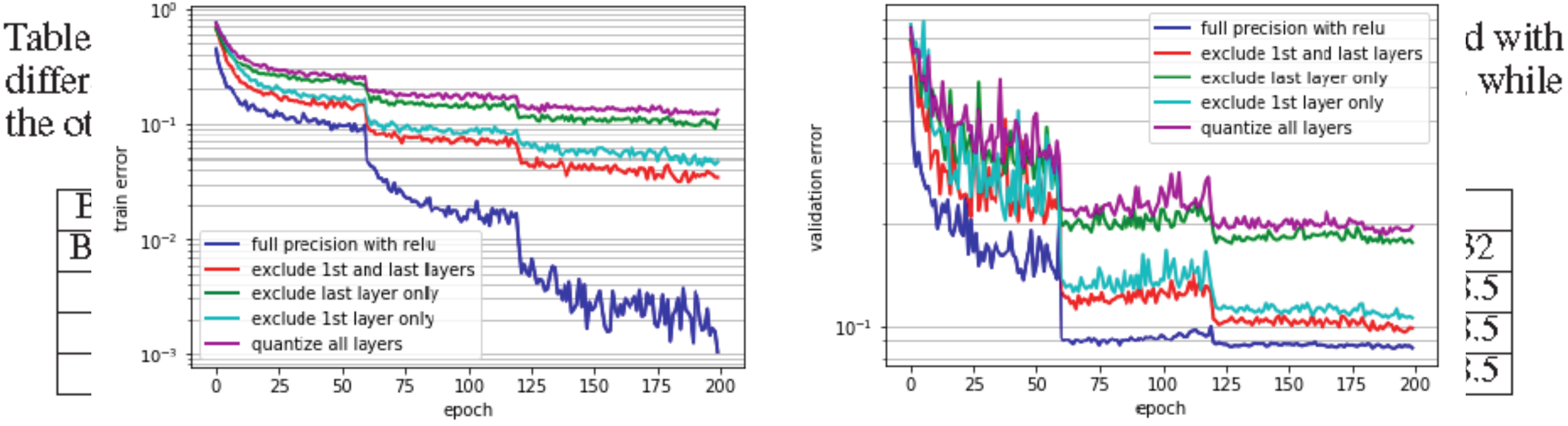
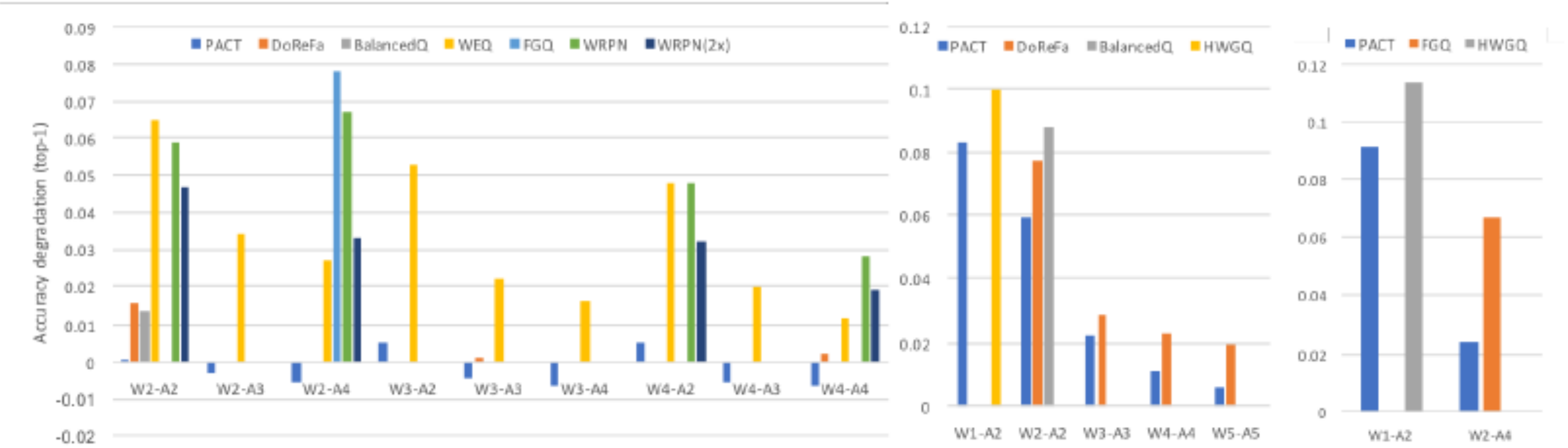


Figure 9: Comparison of accuracy of CIFAR10-ResNet20 with and without quantization of the first and last layers.

EXPERIMENTS

- 实验结果：



(a) AlexNet (b) ResNet18 (c) ResNet50

Figure 5: Comparison of accuracy degradation (Top-1) for (a) AlexNet, (b) ResNet18, and (c) ResNet50.

of accuracy degradation for ResNet18 (left) and ResNet50 (right). The lower the better.

SYSTEM-LEVEL PERFORMANCE GAIN

- 分析了系统上的进步：

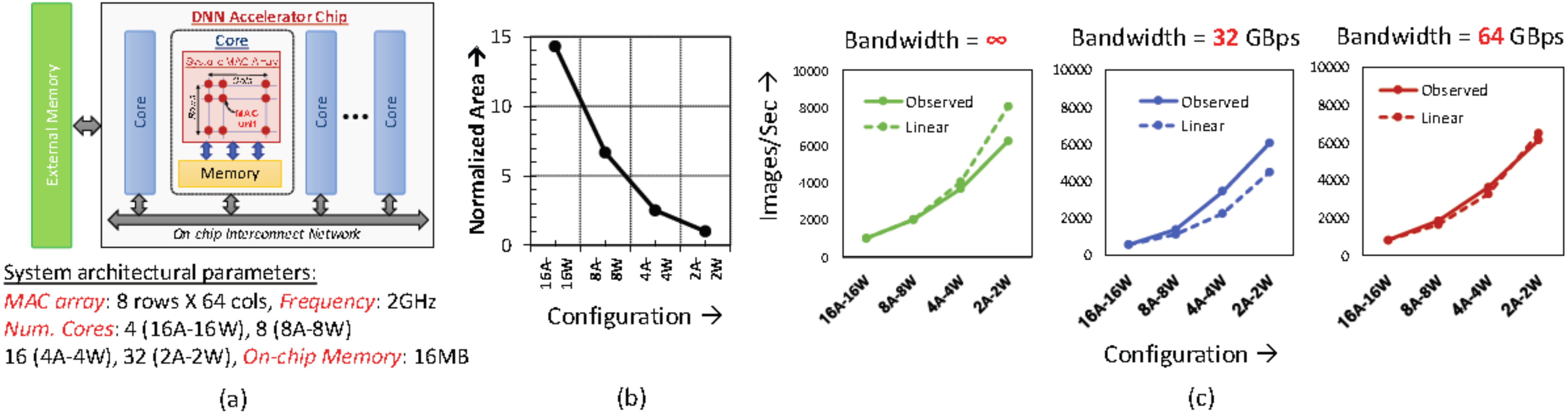


Figure 6: (a)System architecture and parameters, (b) Variation in MAC area with bit-precision and (b) Speedup at different quantizations for inference using ResNet50 DNN