[Supplementary Material] Learning to Quantize Deep Networks by Optimizing Quantization Intervals with Task Loss

1. Top-1 and top-5 accuracy on ImageNet

The top-1 and top-5 accuracy on imageNet with various bit-widths are shown in Table 1 and 2, respectively. We compared our method with other existing methods.

Table 1. Top-1 accuracy (%) on ImageNet. Comparion with the existing methods on ResNet-18, -34 and AlexNet. The 'FP' represents the top-1 accuracy of the full-precision (32/32-bit) network in our implementation.

Method	ResNet-18 (FP: **70.2**)				ResNet-34 (FP: **73.7**)				AlexNet (FP: **61.8**)			
	Bit-width (A/W)											
	5/5	4/4	3/3	2/2	5/5	4/4	3/3	2/2	5/5	4/4	3/3	2/2
QIL (Ours)[†]	**70.4**	**70.1**	**69.2**	**65.7**	**73.7**	**73.7**	**73.1**	**70.6**	**61.9**	**62.0**	**61.3**	**58.1**
LQ-Nets [26]	-	69.3	68.2	64.9	-	-	71.9	69.8	-	-	-	57.4
PACT [4]	69.8	69.2	68.1	64.4	-	-	-	-	55.7	55.7	55.6	55.0
DoReFa-Net [27]	68.4	68.1	67.5	62.6	-	-	-	-	54.9	54.9	55.0	53.6
ABC-Net [17]	65.0	-	61.0	-	68.4	-	66.7	-	-	-	-	-
BalancedQ [28]	-	-	-	59.4	-	-	-	-	-	-	-	55.7
TSQ[†] [25]	-	-	-	-	-	-	-	-	-	-	-	58.0
SYQ[†] [6]	-	-	-	-	-	-	-	-	-	-	-	55.8
Zhuang *et al*. [30]	-	-	-	-	-	-	-	-	-	58.1	-	52.5
WEQ [20]	-	-	-	-	-	-	-	-	-	55.9	54.9	50.6

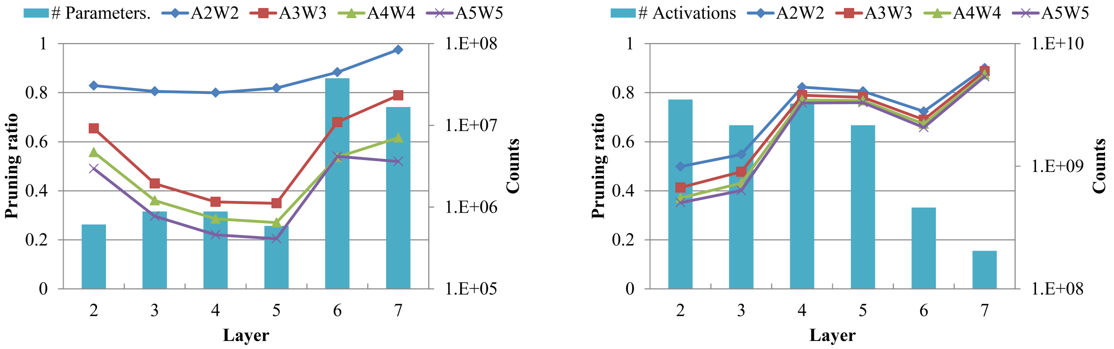
Table 2. Top-5 accuracy (%) on ImageNet. Comparion with the existing methods on ResNet-18, -34 and AlexNet. The 'FP' represents the top-5 accuracy of the full-precision (32/32-bit) network in our implementation.

Method	ResNet-18 (FP: **89.6**)				ResNet-34 (FP: **91.4**)				AlexNet (FP: **83.5**)			
	Bit-width (A/W)											
	5/5	4/4	3/3	2/2	5/5	4/4	3/3	2/2	5/5	4/4	3/3	2/2
QIL (Ours)[†]	**89.6**	**89.5**	**89.0**	**86.8**	**91.6**	**91.7**	**91.3**	**89.9**	**83.8**	**83.8**	**83.3**	**81.0**
LQ-Nets [26]	-	88.8	87.9	85.9	-	-	90.2	89.1	-	-	-	80.1
PACT [4]	89.3	89.0	88.2	85.6	-	-	-	-	77.8	78.0	78.0	77.7
DoReFa-Net [27]	88.3	88.1	87.6	84.4	-	-	-	-	77.9	77.5	77.8	76.8
ABC-Net [17]	85.9	-	83.2	-	88.2	-	87.4	-	-	-	-	-
BalancedQ [28]	-	-	-	82.0	-	-	-	-	-	-	-	78.0
TSQ[†] [25]	-	-	-	-	-	-	-	-	-	-	-	80.5
SYQ[†] [6]	-	-	-	-	-	-	-	-	-	-	-	79.2
Zhuang *et al*. [30]	-	-	-	-	-	-	-	-	-	81.2	-	77.3
WEQ [20]	-	-	-	-	-	-	-	-	-	79.2	78.5	75.0

[†] The 2-bit weights are ternary $\{-1, 0, 1\}$.

2. Layerwise pruning ratio on AlexNet

Fig. 1 shows the pruning ratio of weights and activations for each layer of AlexNet. The 6-th and the 7th layers are fully-connected layers which have larger number of parameters and are more pruned than the convolutional layers.



(a) Weight

(b) Activation

Figure 1. Layerwise pruning ratio of (a) weight and (b) activation for each layer of AlexNet. The bar graph shows the number of weights or activations for each layer on a log scale.