HAQ: Hardware-Aware Automated Quantization with Mixed Precision

		Edge Accelerator						Cloud Accelerator					
		MobileNet-V1			MobileNet-V2			MobileNet-V1			MobileNet-V2		
	Bitwidths	Acc.-1	Acc.-5	Latency	Acc.-1	Acc.-5	Latency	Acc.-1	Acc.-5	Latency	Acc.-1	Acc.-5	Latency
PACT [3]	4 bits	62.44	84.19	45.45 ms	61.39	83.72	52.15 ms	62.44	84.19	57.49 ms	61.39	83.72	74.46 ms
Ours	*flexible*	**67.40**	**87.90**	45.51 ms	**66.99**	**87.33**	52.12 ms	**65.33**	**86.60**	57.40 ms	**67.01**	**87.46**	73.97 ms
PACT [3]	5 bits	67.00	87.65	57.75 ms	68.84	88.58	66.94 ms	67.00	87.65	77.52 ms	68.84	88.58	99.43 ms
Ours	*flexible*	**70.58**	**89.77**	57.70 ms	**70.90**	**89.91**	66.92 ms	**69.97**	**89.37**	77.49 ms	**69.45**	**88.94**	99.07 ms
PACT [3]	6 bits	70.46	89.59	70.67 ms	71.25	90.00	82.49 ms	70.46	89.59	99.86 ms	71.25	90.00	127.07 ms
Ours	*flexible*	**71.20**	**90.19**	70.35 ms	**71.89**	**90.36**	82.34 ms	**71.20**	**90.08**	99.66 ms	**71.85**	**90.24**	127.03 ms
Original	8 bits	70.82	89.85	96.20 ms	71.81	90.25	115.84 ms	70.82	89.85	151.09 ms	71.81	90.25	189.82 ms

Table 3: Latency-constrained quantization on BISMO (edge accelerator and cloud accelerator) on ImageNet. Our framework can reduce the latency by **1.4**$\times$ to **1.95**$\times$ with negligible loss of accuracy compared with the fixed bitwidth (8 bits) quantization.

	Weights	Activations	Acc.-1	Acc.-5	Latency
PACT [3]	4 bits	4 bits	62.44	84.19	7.86 ms
Ours	*flexible*	*flexible*	**67.45**	**87.85**	7.86 ms
PACT [3]	6 bits	4 bits	67.51	87.84	11.10 ms
Ours	*flexible*	*flexible*	**70.40**	**89.69**	11.09 ms
PACT [3]	6 bits	6 bits	70.46	89.59	19.99 ms
Ours	*flexible*	*flexible*	**70.90**	**89.95**	19.98 ms
Original	8 bits	8 bits	70.82	89.85	20.08 ms

Table 4: Latency-constrained quantization on BitFusion (MobileNet-V1 on ImageNet). Our framework can reduce the latency by **2**$\times$ with almost no loss of accuracy compared with the fixed bitwidth (8 bits) quantization.

	Weights	Activations	Acc.-1	Acc.-5	Energy
PACT [3]	4 bits	4 bits	62.44	84.19	13.47 mJ
Ours	*flexible*	*flexible*	**64.78**	**85.85**	13.69 mJ
PACT [3]	6 bits	4 bits	67.51	87.84	16.57 mJ
Ours	*flexible*	*flexible*	**70.37**	**89.40**	16.30 mJ
PACT [3]	6 bits	6 bits	70.46	89.59	26.80 mJ
Ours	*flexible*	*flexible*	**70.90**	**89.73**	26.67 mJ
Original	8 bits	8 bits	70.82	89.95	31.03 mJ

Table 5: Energy-constrained quantization on BitFusion (MobileNet-V1 on ImageNet). Our framework reduces the power consumption by **2**$\times$ with nearly no loss of accuracy compared with the fixed bitwidth quantization.

		MobileNet-V1			MobileNet-V2			ResNet-50		
	Weights	Acc.-1	Acc.-5	Model Size	Acc.-1	Acc.-5	Model Size	Acc.-1	Acc.-5	Model Size
Han *et al.* [9]	2 bits	37.62	64.31	1.09 MB	58.07	81.24	0.96 MB	68.95	88.68	6.32 MB
Ours	*flexible*	**57.14**	**81.87**	1.09 MB	**66.75**	**87.32**	0.95 MB	**70.63**	**89.93**	6.30 MB
Han *et al.* [9]	3 bits	65.93	86.85	1.60 MB	68.00	87.96	1.38 MB	75.10	92.33	9.36 MB
Ours	*flexible*	**67.66**	**88.21**	1.58 MB	**70.90**	**89.76**	1.38 MB	**75.30**	**92.45**	9.22 MB
Han *et al.* [9]	4 bits	71.14	89.84	2.10 MB	71.24	89.93	1.79 MB	**76.15**	92.88	12.40 MB
Ours	*flexible*	**71.74**	**90.36**	2.07 MB	**71.47**	**90.23**	1.79 MB	76.14	**92.89**	12.14 MB
Original	32 bits	70.90	89.90	16.14 MB	71.87	90.32	13.37 MB	76.15	92.86	97.49 MB

Table 6: Model size-constrained quantization on ImageNet. Compared with Deep Compression [8], our framework achieves higher accuracy under similar model size (especially under high compression ratio).

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

Table 6. Comparison of 1-bit quantized models on CIFAR-10.

Model	Method	Bit-Width (W/A)	Accuracy (%)
VGG-Small	FP	32/32	91.65
	BNN [16]	1/1	89.90
	XNOR [33]	1/1	89.80
	Ours	1/1	**91.72**
ResNet-20	FP	32/32	90.78
	DoReFa [43]	1/1	79.30
	Ours	1/1	**84.11**
	DoReFa [43]	1/32	90.00
	LQ-Net [41]	1/32	90.10
	Ours	1/32	**90.24**

Table 7. Comparison of different quantized models on ImageNet.

Model	Method	Bit-Width (W/A)	Accuracy (%)
ResNet-18	FP	32/32	69.90
	BWN [33]	1/32	60.80
	HWGQ [5]	1/32	61.30
	TWN [23]	2/32	61.80
	Ours	1/32	**63.71**
	PACT [7]	2/2	64.40
	LQ-Net [41]	2/2	64.90
	Ours	2/2	**65.17**
	ABC-Net [25]	3/3	61.00
	PACT [7]	3/3	68.10
	LQ-Net [41]	3/3	68.20
	Ours	3/3	**68.66**
	BCGD [40]	4/4	67.36†
	Ours	4/4	**69.56**†
ResNet-34	FP	32/32	73.80
	LQ-Net [41]	2/2	69.80
	Ours	2/2	**70.02**
	ABC-Net [25]	3/3	66.70
	LQ-Net [41]	3/3	71.90
	Ours	3/3	**72.54**
	BCGD [40]	4/4	70.81†
	Ours	4/4	**72.76**†
Mobile-NetV2	FP	32/32	71.87
	PACT [7, 36]	4/4	61.40
	Ours	4/4	**64.80**

* The † represents the results of full quantization for activations and weights across all convolution layers.

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

Algorithm	Bits	Top-1 Bef.	Top-1 Aft.	Top-1 ↓	Top-5 Bef.	Top-5 Aft.	Top-5 ↓	Mult.↓	Add.↓	Weight↓
Pruning [Han *et al.*, 2015]	-	57.2%	57.2%	0.0%	80.3%	80.3%	0.0%	3×	3×	3×
Sparsification [Figurnov *et al.*, 2016]	-	-	-	-	80.4%	70.5%	9.9%	4.4×	4.4×	-
	-	-	-	-		74.3%	6.1%	3.5×	3.5×	-
	-	-	-	-		78.1%	2.3%	2.1×	2.1×	-
Low-rank [Tai *et al.*, 2015]	-	-	-	-	80.0%	79.6%	0.4%	5.27×	5.27×	5.00×
Decomposition [Kim *et al.*, 2015]	-	-	-	-	80.0%	78.3%	1.7%	2.67×	2.67×	5.46×
EEC [Yang *et al.*, 2017]	-	-	-	-	80.0%	79.5%	0.5%	6.66×	6.66×	11×
NISP [Yu *et al.*, 2018]	-	-	-	-	80.0%	80.0%	0.0%	2.5×	2.5×	2.1×
BWN* [Courbariaux *et al.*, 2015]	1	56.6%	29.9%	26.7%	80.0%	52.7%	37.3%	1656×	1.0×	30.6×
ABC* [Lin *et al.*, 2017]	2	56.6%	52.4%	4.2%	80.0%	76.3%	3.7%	828×	0.49×	15.81×
	3		54.0%	2.6%		77.7%	2.3%	552×	0.32×	10.54×
	4		53.5%	3.1%		77.2%	2.8%	414×	0.24×	7.90×
	5		55.9%	0.7%		79.2%	0.8%	331×	0.19×	6.32×
KCNN	1	56.6%	40.4%	16.2%	80.0%	65.3%	14.7%	1656×	1.0×	30.6×
	2		53.7%	2.9%		77.2%	2.8%	828×	0.49×	15.81×
	3		55.2%	1.4%		78.6%	1.4%	552×	0.32×	10.54×
	4		56.4%	0.2%		79.6%	0.4%	414×	0.24×	7.90×
	5		56.4%	0.2%		79.5%	0.5%	331×	0.19×	6.32×
KCNN + Low-rank	1	56.6%	37.9%	18.7%	80.0%	62.2%	17.8%	1434×	5.27×	149.6×
	2		51.8%	4.8%		75.7%	4.3%	717×	2.62×	74.8×
	3		53.6%	3.0%		77.1%	2.9%	478×	1.74×	49.8×
	4		55.6%	1.0%		78.8%	1.2%	358×	1.30×	37.4×
	5		56.2%	0.4%		79.5%	0.5%	286×	1.03×	29.9×

Table 1: The comparison between our proposed KCNN and previous methods on AlexNet.
*The BWN and ABC are realized by us.

Algorithm	Bits	Top-1 Bef.	Top-1 Aft.	Top-1 ↓	Top-5 Bef.	Top-5 Aft.	Top-5 ↓	Mult.↓	Add.↓	Weight↓
BWN [Courbariaux *et al.*, 2015]	1	69.3%	60.8%	8.5%	89.2%	83.0%	6.2%	730×	1×	31.5×
TWN [Li *et al.*, 2016]	2	-	61.8%	-	-	84.2%	-	730×	1×	31.5×
ABC [Lin *et al.*, 2017]	1	69.3%	62.8%	6.5%	89.2%	84.4%	4.8%	730×	1×	31.5×
	2		63.7%	5.6%		85.2%	4.0%	365×	0.49×	15.7×
	3		66.2%	3.1%		86.7%	2.5%	243×	0.32×	10.5×
	5		68.3%	1.0%		87.9%	1.3%	146×	0.19×	6.3×
KCNN	1	69.2%	61.7%	7.5%	89.0%	84.2%	4.8%	730×	1×	31.5×
	2		66.5%	2.7%		87.4%	1.6%	365×	0.49×	15.7×
	3		67.6%	1.6%		88.1%	0.9%	243×	0.32×	10.5×
	4		68.3%	0.9%		88.5%	0.5%	182×	0.24×	7.8×
	5		68.7%	0.5%		88.7%	0.3%	146×	0.19×	6.3×

Table 2: The comparison between our proposed KCNN and previous methods on ResNet-18.

Table I: *Quantization results of ResNet20 on Cifar-10. We abbreviate quantization bits used for weights as "w-bits," activations as "a-bits," testing accuracy as "Acc," and compression ratio of weights/activations as "W-Comp/A-Comp." Furthermore, we show results without using Hessian information ("Direct"), as well as other state-of-the-art methods [43], [2], [40]. In particular, we compare with the recent DNAS approach of [36]. Our method achieves similar testing performance with significantly higher compression (especially in activations). Here "MP" refers to mixed-precision quantization, where we report the lowest bits used for weights and activations. Also note that [43], [2], [40] use 8-bit for first and last layers. The exact per-layer configuration for mixed-precision quantization of* HAWQ *is presented in appendix.*

Quantization	w-bits	a-bits	Acc	W-Comp	A-Comp
Baseline	32	32	92.37	1.00×	1.00×
Dorefa [43]	2	2	88.20	16.00×	16.00×
Dorefa [43]	3	3	89.90	10.67×	10.67×
PACT [2]	2	2	89.70	16.00×	16.00×
PACT [2]	3	3	91.10	10.67×	10.67×
LQ-Nets [40]	2	2	90.20	16.00×	16.00×
LQ-Nets [40]	3	3	91.60	10.67×	10.67×
LQ-Nets [40]	2	32	91.80	16.00×	1.00×
LQ-Nets [40]	3	32	92.00	10.67×	1.00×
DNAS [36]	1 MP	32	92.00	16.60×	1.00×
DNAS [36]	1 MP	32	**92.72**	11.60×	1.00×
Direct	2 MP	4	90.34	16.00×	8.00×
HAWQ	2 MP	4	**92.22**	13.11×	8.00×

Table III: *Quantization results of ResNet50 on ImageNet. We show results of state-of-the-art methods [43], [2], [40], [8]. In particular, we also compare with the recent AutoML approach of [35]. Compared to [35], we achieve higher compression ratio with higher testing accuracy. Also note that [43], [2], [40] use 8-bit for first and last layers.*

Method	w-bits	a-bits	Top-1	W-Comp	Size(MB)
Baseline	32	32	77.39	1.00×	97.8
Dorefa [43]	2	2	67.10	16.00×	6.11
Dorefa [43]	3	3	69.90	10.67×	9.17
PACT [2]	2	2	72.20	16.00×	6.11
PACT [2]	3	3	75.30	10.67×	9.17
LQ-Nets [40]	3	3	74.20	10.67×	9.17
Deep Comp. [8]	3	MP	75.10	10.41×	9.36
HAQ [35]	MP	MP	75.30	10.57×	9.22
HAWQ	2 MP	4 MP	**75.48**	12.28×	**7.96**

Table IV: *Quantization results of SqueezeNext on ImageNet. We show a case where* HAWQ *is used to achieved uniform quantization to 8 bits for both weights and activations, with an accuracy similar to ResNet18. We also show a case with mixed precision, where we compress SqueezeNext to a model with just 1MB size with only 1.36% accuracy degradataion. Furthermore, we compare* HAWQ *with direct quantization method without using Hessian ("Direct").*

Method	w-bits	a-bits	Top-1	W-Comp	Size(MB)
Baseline	32	32	69.38	1.00×	10.1
ResNet18 [27]	32	32	69.76	1.00×	44.7
HAWQ	8	8	**69.34**	4.00×	2.53
Direct	3 MP	8	**65.39**	9.04×	1.12
HAWQ	3 MP	8	**68.02**	9.25×	**1.09**

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

Table 1: *Quantization results of Inception-V3 on ImageNet. We abbreviate quantization bits used for weights as "w-bits," quantization bits used for activations as "a-bits," top-1 testing accuracy as "Top-1," and weight compression ratio as "W-Comp." Furthermore, we compare* HAWQ-V2 *with direct quantization method without using Hessian ("Direct") and Integer-Only [11]. Here "MP" refers to mixed-precision quantization. Compared to [11, 18], we achieve higher compression ratio with higher testing accuracy.*

Method	w-bits	a-bits	Top-1	W-Comp	Size(MB)
Baseline	32	32	77.45	1.00×	91.2
Integer-Only [11]	8	8	75.40	4.00×	22.8
Integer-Only [11]	7	7	75.00	4.57×	20.0
RVQuant [18]	3 MP	3 MP	74.14	10.67×	8.55
Direct	2 MP	4 MP	69.76	15.88×	5.74
HAWQ [7]	2 MP	4 MP	75.52	12.04×	**7.57**
HAWQ-V2	2 MP	4 MP	**75.68**	12.04×	**7.57**

We also show HAWQ-V2 results on ResNet50 [10], and compare HAWQ-V2 with other popular quantization methods [5, 7, 9, 22, 28, 30] in Table 2. It should be noted that [5, 9, 28, 30] followed traditional quantization rules which set the precision for the first and last layer to 8-bit, and quantized other layers to an identical precision. Both [7, 22] are mixed-precision quantization methods. Also, [22] uses reinforcement learning methods to search for a good precision setting, while HAWQ uses second-order information to guide the precision selection as well as the block-wise fine-tuning. HAWQ achieves the state-of-the-art accuracy 75.48% with a 7.96MB model size. Keeping model size the same, HAWQ-V2 can achieve 75.76% accuracy without any heuristic knowledge and manual efforts.

Table 2: *Quantization results of ResNet50 on ImageNet. We show results of state-of-the-art methods [5, 9, 28, 30]. We also compare with the recent AutoML approach of [22]. Compared to [22], we achieve higher compression ratio with higher testing accuracy. Also note that [5, 28, 30] use 8-bit for first and last layers.*

Method	w-bits	a-bits	Top-1	W-Comp	Size(MB)
Baseline	32	32	77.39	1.00×	97.8
Dorefa [30]	2	2	67.10	16.00×	6.11
Dorefa [30]	3	3	69.90	10.67×	9.17
PACT [5]	2	2	72.20	16.00×	6.11
PACT [5]	3	3	75.30	10.67×	9.17
LQ-Nets [28]	3	3	74.20	10.67×	9.17
Deep Comp. [9]	3	MP	75.10	10.41×	9.36
HAQ [22]	MP	MP	75.30	10.57×	9.22
HAWQ [7]	2 MP	4 MP	75.48	12.28×	7.96
HAWQ-V2	2 MP	4 MP	**75.76**	12.24×	**7.99**

Table 3: *Quantization results of SqueezeNext on ImageNet. We first show results of direct quantization method without using Hessian ("Direct"). Then we compare* HAWQ-V2 *with* HAWQ, *which can compress SqueezeNext to a model with an unprecedented 1MB model size with only 1.36% top-1 accuracy drop. By applying* HAWQ-V2 *on SqueezeNext, we can achieve even better accuracy* 68.38% *with even smaller model size than* HAWQ.

Method	w-bits	a-bits	Top-1	W-Comp	Size(MB)
Baseline	32	32	69.38	1.00×	10.1
Direct	3 MP	8	65.39	9.04×	1.12
HAWQ [7]	3 MP	8	68.02	9.26×	**1.09**
HAWQ-V2	3 MP	8	**68.38**	9.40×	**1.07**

object categories. RetinaNet [14] is a single stage detector that can achieve state-of-the-art mAP[4] with a very simple network architecture, and it only contains hardware-friendly operations such as convolutions and additions. As shown in Table 4, we use the pretrained RetinaNet with ResNet50 backbone as our baseline model, which can achieve 35.6 mAP with 145MB model size. We first show the result of direct quantization where no Hessian information is used. Even with quantization-aware fine-tuning and channel-wise quantization of weights, directly quantizing weights and activations in RetinaNet to 4-bit causes a significant 4.1 mAP degradation. FQN [13] is a recently proposed quantization method which reduces this accuracy gap to 3.1 mAP with the same compression ratio as Direct method. Using HAWQ-V2 on mixed-precision weight quantization with uniform 4-bit activations can achieve a state-of-the-art performance of 34.1 mAP, which is 1.6 mAP higher than [13] with an even smaller model size.

Table 4: *Quantization results of RetinaNet on Microsoft COCO 2017. We show results of direct quantization, as well as a state-of-the-art quantization method for object detection [13]. With the same model size,* HAWQ-V2 *can outperform previous quantization results by a large margin. We also show that* HAWQ-V2 *with mixed-precision activations can achieve even better mAP, with a slightly lower activation compression ratio.*

Method	w-bits	a-bits	mAP	W-Comp	A-Comp	Size(MB)
Baseline	32	32	35.6	1.00×	1.00×	145
Direct	4	4	31.5	8.00×	8.00×	18.13
FQN [13]	4	4	32.5	8.00×	8.00×	18.13
HAWQ-V2	3 MP	4	**34.1**	8.10×	8.00×	17.90
HAWQ-V2	3 MP	4 MP	**34.4**	8.10×	7.62×	17.90
HAWQ-V2	3 MP	6	**34.8**	8.10×	5.33×	17.90

AdaBits: Neural Network Quantization with Adaptive Bit-Widths

Scheme	Individual Quantization (SAT)				Adaptive Bit-widths			BitOPs
	Name	Bit-width	Size	Top-1 Acc.	Name	Size	Top-1 Acc.	
Original	MobileNet V1	8 bit	4.10 MB	72.6	AB-MobileNet V1 [8, 6, 5, 4] bits	FP	72.4 (-0.2)	36.40 B
	MobileNet V1	6 bit	3.34 MB	72.3			72.4 (0.1)	20.81 B
	MobileNet V1	5 bit	2.96 MB	71.9			72.1 (0.2)	14.68 B
	MobileNet V1	4 bit	2.58 MB	71.3			71.1 (-0.2)	9.67 B
	MobileNet V2	8 bit	3.44 MB	72.5	AB-MobileNet V2 [8, 6, 5, 4] bits	FP	72.6 (0.1)	19.25 B
	MobileNet V2	6 bit	2.92 MB	72.3			72.4 (0.1)	11.17 B
	MobileNet V2	5 bit	2.66 MB	72.0			72.1 (0.1)	7.99 B
	MobileNet V2	4 bit	2.40 MB	71.1			70.8 (-0.3)	5.39 B
	ResNet50	4 bit	13.34 MB	76.3	AB-ResNet50 [4, 3, 2] bits	FP	76.1 (-0.2)	71.81 B
	ResNet50	3 bit	10.55 MB	75.9			75.8 (-0.1)	43.75 B
	ResNet50	2 bit	7.75 MB	73.3			73.2 (-0.1)	23.71 B
Modified	MobileNet V1	8 bit	4.10 MB	72.6	AB-MobileNet V1 [8, 6, 5, 4] bits	4.35 MB	72.3 (-0.3)	36.40 B
	MobileNet V1	6 bit	3.34 MB	72.4			72.3 (-0.1)	20.81 B
	MobileNet V1	5 bit	2.96 MB	72.2			72.0 (-0.2)	14.68 B
	MobileNet V1	4 bit	2.58 MB	70.5			70.4 (-0.1)	9.67 B
	MobileNet V2	8 bit	3.44 MB	72.7	AB-MobileNet V2 [8, 6, 5, 4] bits	3.83 MB	72.3 (-0.4)	19.25 B
	MobileNet V2	6 bit	2.92 MB	72.5			72.3 (-0.2)	11.17 B
	MobileNet V2	5 bit	2.66 MB	72.1			72.0 (-0.1)	7.99 B
	MobileNet V2	4 bit	2.40 MB	70.3			70.3 (0.0)	5.39 B

Table 4. Comparison between individual quantization and AdaBits quantization for top-1 validation accuracy (%) of MobileNet V1/V2 and ResNet50 on ImageNet. Note that we use two quantization schemes to compare our AdaBits with SAT baseline models where "original" denotes the original DoReFa scheme and "modified" denote the modified scheme in Eq. (3) which enables producing weights for lower bit-width from the 8-bit model. "FP" denotes the full-precision models is needed to recover weights in different bit-widths.

Towards Efficient Training for Neural Network Quantization

Table 1: Comparison of quantization techniques with both weights and activation quantized.

		MobileNet-V1		MobileNet-V2	
Quant. Method	Bit-widths	Acc.-1	Acc.-5	Acc.-1	Acc.-5
PACT	4bits	70.3	89.2	70.4	89.4
HAQ	flexible	67.40	87.90	66.99	87.33
SAT (Ours)	4bits	**71.3**	**89.9**	**71.1**	**90.0**
PACT	5bits	71.1	89.6	71.2	89.8
HAQ	flexible	70.58	89.77	70.90	89.91
SAT (Ours)	5bits	**71.9**	**90.3**	**72.0**	**90.4**
PACT	6bits	71.2	89.2	71.5	90.0
HAQ	flexible	71.20	90.19	71.89	90.36
SAT (Ours)	6bits	**72.3**	**90.4**	**72.3**	**90.6**
PACT	8bits	71.3	89.7	71.7	89.9
HAQ	flexible	70.82	89.85	71.81	90.25
SAT (Ours)	8bits	**72.6**	**90.7**	**72.5**	**90.7**
PACT	FP	72.1	90.2	72.1	90.5
SAT (Ours)	FP	71.7	90.2	71.8	90.2

Table 2: Comparison of quantization techniques with only weights quantized.

		MobileNet-V1		MobileNet-V2	
Quant. Method	Weights	Acc.-1	Acc.-5	Acc.-1	Acc.-5
Deep Compression	2bits	37.62	64.31	58.07	81.24
HAQ	flexible	57.14	81.87	**66.75**	**87.32**
SAT (Ours)	2bits	**66.3**	**86.8**	**66.8**	87.2
Deep Compression	3bits	65.93	86.85	68.00	87.96
HAQ	flexible	67.66	88.21	70.90	89.76
SAT (Ours)	3bits	**70.7**	**89.5**	**71.1**	**89.9**
Deep Compression	4bits	71.14	89.84	71.24	89.93
HAQ	flexible	71.74	**90.36**	71.47	90.23
SAT (Ours)	4bits	**72.1**	90.2	**72.1**	**90.6**
Deep Compression	FP	70.90	89.90	71.87	90.32
HAQ	FP	70.90	89.90	71.87	90.32
SAT (Ours)	FP	71.7	90.2	71.8	90.2

Both Quantization				Weight-Only Quantization			
Quant. Method†	Bit-widths	Acc.-1	Acc.-5	Quant. Method†	Weights	Acc.-1	Acc.-5
PACT	2bits	72.2	90.5	DeepCompression	2bits	68.95	88.68
LQNet	2bits	71.5	90.3	LQNet	2bits	75.1	92.3
LSQ	2bits	73.7	91.5	HAQ	flexible	70.63	89.93
SAT (Ours)	2bits	**74.1**	**91.7**	SAT (Ours)	2bits	**75.3**	**92.4**
PACT	3bits	75.3	92.6	DeepCompression	3bits	75.10	92.33
LQNet	3bits	74.2	91.6	LQNet	3bits	NA	NA
LSQ	3bits	75.8	92.7	HAQ	flexible	75.30	92.45
SAT (Ours)	3bits	**76.6**	**93.1**	SAT (Ours)	3bits	**76.3**	**93.0**
PACT	4bits	76.5	93.2	DeepCompression	4bits	76.15	92.88
LQNet	4bits	75.1	92.4	LQNet	4bits	**76.4**	**93.1**
LSQ	4bits	76.7	93.2	HAQ	flexible	76.14	92.89
SAT (Ours)	4bits	**76.9**	**93.3**	SAT (Ours)	4bits	**76.4**	93.0
PACT	FP	76.9	93.1	DeepCompression	FP	76.15	92.86
LQNet	FP	76.4	93.2	LQNet	FP	76.4	93.2
LSQ	FP	76.9	93.4	HAQ	FP	76.15	92.86
SAT (Ours)	FP	75.9	92.5	SAT (Ours)	FP	75.9	92.5

* PACT and SAT use full pre-activation ResNet, LSQ and HAQ use vanilla ResNet, and LQNet uses vanilla ResNet without convolution operation in shortcut (type-A shortcut).

† PACT and LQNet use full-precision for the first and last layers, LSQ and SAT use 8bit for both layers, and HAQ uses 8bit for the first layer.

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

Table 1. Top-1 accuracy (%) on ImageNet. Comparion with the existing methods on ResNet-18, -34 and AlexNet. The 'FP' represents the full-precision (32/32-bit) accuracy 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. The top-1 accuracy (%) of low bit-width networks on ResNet-18 with direct and progressive finetuning. The 5/5-bit network was finetuned from full-precision network.

Initialization	Bit-width (A/W)				
	32/32	5/5	4/4	3/3	2/2
Direct	70.2	70.4	69.9	68.7	56.0
Progressive		-	70.1	69.2	65.7

Table 3. Joint training vs. Quantizer only. The top-1 accuracy (%) with ResNet-18.

Initialization	Bit-width (A/W)				
	32/32	5/5	4/4	3/3	2/2
Joint training	70.2	70.4	70.1	69.2	65.7
Quantizer only		69.4	68.0	62.0	20.9

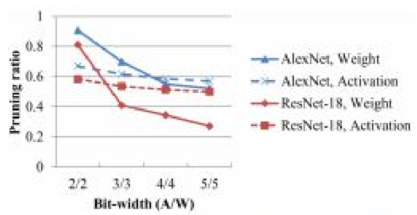

Figure 3. Average pruning ratio of weights and activations on AlexNet and ResNet-18 with various bit-widths

or we can optimize only the quantizers while keeping the weight parameters fixed. Table 3 shows the top-1 accuracy with ResNet-18 network on the both cases. Both the cases utilize the progressive finetuning. The joint training of quantizer and weights works better than training

AUTOQ: AUTOMATED KERNEL-WISE NEURAL NETWORK QUANTIZATION*

Table 3: Network Quantization by AutoQ (A-QBN: the average QBN of activations; W-QBN: the average QBN of weights; LAT: inference latency).

model	scheme	resource-constrained					accuracy-guaranteed				
		top-1 err (%)	top-5 err(%)	A-QBN (bit)	W-QBN (bit)	LAT (ms)	top-1 err (%)	top-5 err(%)	A-QBN (bit)	W-QBN (bit)	LAT (ms)
ResNet-18	network-wise	32.7	12.32	4	4	296.8	32.7	12.32	4	4	296.8
	layer-wise	31.8	11.92	3.32	4.63	290.9	32.5	11.90	3.37	3.65	189.6
	kernel-wise	**30.22**	**11.62**	4.12	3.32	286.3	32.6	11.82	**3.02**	**2.19**	125.3
	original	30.10	11.62	16	16	1163	30.10	11.62	16	16	1163
ResNet-50	network-wise	27.57	9.02	4	4	616.3	27.57	9.02	4	4	616.3
	layer-wise	26.79	8.32	4.23	3.51	612.3	27.49	9.15	4.02	3.12	486.4
	kernel-wise	**25.53**	**7.92**	3.93	4.02	610.3	27.53	9.12	**3.07**	**2.21**	327.3
	original	25.20	7.82	16	16	2357	25.20	7.82	16	16	2357
SqueezeNetV1	network-wise	45.67	23.12	4	4	43.1	45.67	23.12	4	4	43.1
	layer-wise	44.89	21.14	3.56	4.27	42.1	45.63	23.04	3.95	3.28	25.5
	kernel-wise	**43.51**	**20.89**	4.05	3.76	41.6	45.34	23.02	**3.29**	**2.32**	12.5
	original	43.10	20.5	16	16	127.3	43.10	20.5	16	16	127.3
MobileNetV2	network-wise	31.75	11.67	4	4	37.4	31.35	11.67	4	4	37.4
	layer-wise	30.98	10.57	3.57	4.22	36.9	31.34	10.57	3.92	3.21	23.9
	kernel-wise	**29.20**	**9.67**	4.14	3.67	36.1	31.32	11.32	**3.13**	**2.26**	10.2
	original	28.90	9.37	16	16	123.6	28.90	9.37	16	16	123.6

MIXED PRECISION DNNS: ALL YOU NEED IS A GOOD PARAMETRIZATION

Table 3: Homogeneous vs. heterogeneous quantization of ResNet-20 on CIFAR-10.

	Bitwidth Weight/Activ.	q_{max} Weight/Activ.	Size Weight/Activ.(max)/Activ.(sum)	**Uniform quant.** Validation error	**Power-of-two quant.** Validation error
Baseline	32bit/32bit	–	1048KB/64KB/736KB	7.29%	
Fixed	2bit/32bit	fixed/–	65.5KB/64KB/736KB	10.81%	8.99%
TQT (Jain et al., 2019)	2bit/32bit	learned/ –	65.5KB/64KB/736KB	9.47%	8.79%
Ours (w/ constr. (8a))	learned/32bit	learned/-	70KB/64KB/736KB	8.59%	8.53%
Fixed	2bit/4bit	fixed/fixed	65.5KB/8KB/92KB	11.30%	11.62%
TQT (Jain et al., 2019)	2bit/4bit	learned/learned	65.5KB/8KB/92KB	9.62%	11.29%
Ours (w/ constr. (8a) and (8b))	learned/learned	learned/learned	70KB/ – /92KB	9.38%	11.29%
Ours (w/ constr. (8a) and (8c))	learned/learned	learned/learned	70KB/8KB/ –	8.58%	11.23%

Table 4: Homogeneous vs. heterogeneous quantization of MobileNetV2 and ResNet-18 on ImageNet.

	Bitwidth Weight/Activ.	q_{max} Weight/Activ.	**MobileNetV2** Size Weight/Activ(max)	**MobileNetV2** Validation Error	**ResNet-18** Size Weight/Activ(max)	**ResNet-18** Validation Error
Baseline	32bit/32bit	–	13.23MB/4.59MB	29.82%	44.56MB/3.04MB	29.72%
Fixed	4bit/4bit	fixed/fixed	1.65MB/0.57MB	36.27%	5.57MB/0.38MB	34.15%
TQT (Jain et al., 2019)	4bit/4bit	learned/learned	1.65MB/0.57MB	32.21%	5.57MB/0.38MB	30.49%
Ours (w/ constr. (8a) and (8c))	learned/learned	learned/learned	1.55MB/0.57MB	30.26%	5.40MB/0.38MB	29.92%
Ours (w/o constr.)	learned/learned	learned/learned	3.14MB/1.58MB	29.41%	10.50MB/1.05MB	29.34%