Supplementary Material: Pruning as a Binarization Technique

Lukas Frickenstein[1], Pierpaolo Mori[1,3], Shambhavi Balamuthu Sampath[1], Moritz Thoma[1],
Nael Fasfous[1], Manoj Rohit Vemparala[1], Alexander Frickenstein[1], Christian Unger[1],
Claudio Passerone[3], Walter Stechele[2]
[1]BMW Group, Munich, Germany; [2]Technical University of Munich, Munich, Germany; [3]Politecnico Di Torino, Turin, Italy
{<firstname>.<lastname>}[1]@bmw.de,[2]@tum.de,[3]@polito.it,

S1. Latent-Weight-Free Training of MBNs

One of the contributions of this work is to bring the latent-weight-free optimizer (BOP [4]) to the multi-bit network (MBN) domain. To better understand the effect of BOP on MBNs and its influence on the Top-1 accuracy, Tab. 1 shows an ablation study for different network architectures and datasets trained with different training hyper-parameters. The ablation study is performed across two different optimizers (ADAM [5] and AMSGrad [8]) training the network parameters for batch norm ($\theta_{\mathtt{BN}}$) and operation-oriented scaling factors ($\theta_{\mathtt{SF}}$), while the weight parameters ($\theta_{\mathtt{W}}$) are trained with BOP [4]. For the latter, we consider different values for the adaptivity rate φ and the threshold τ steering the learning process of the weights, while the learning rate η is considered in the optimizer scope of ADAM or AMSGrad. The learnings from this ablation study led to the training hyper-parameters chosen for the networks presented in the main paper. For CIFAR-10 [6], 50K train and 10K test images (32×32 pixels) are used to train and evaluate the multi-bit configurations of ResNet-20/56 [3]. ImageNet [9] consists of $\sim$ 1.28M train and 50K validation images (256×256 pixels), where multi-bit configurations of ResNet-18 [3] are trained and evaluated. The presented network architectures for CIFAR-10 are trained for 500 epochs with varying initial learning rates $\eta \in \{1\text{e-}2; 1\text{e-}3\}$, adaptivity rates $\varphi \in \{1\text{e-}3; 1\text{e-}4\}$ and thresholds $\tau \in \{1\text{e-}6; 1\text{e-}7; 1\text{e-}8\}$. η and φ are decayed by 0.1 every 100 epochs (step-wise). For ImageNet experiments, the network configurations are trained for 100 epochs, where we vary the threshold $\tau \in \{1\text{e-}7; 1\text{e-}8\}$ and the initial adaptivity rate $\varphi = 1\text{e-}4$ is decayed linearly to the final $\varphi \in \{1\text{e-}6; 1\text{e-}8\}$ for weight training. To update the remaining network parameters ($\theta_{\mathtt{BN}}$, $\theta_{\mathtt{SF}}$), we explore the effect of the optimizers ADAM and AMSGrad, where the initial learning rate is $\eta \in \{1\text{e-}3; 2.5\text{e-}3\}$ decayed linearly down to $\eta \in \{5\text{e-}6; 5\text{e-}8\}$. Note that all multi-bit network configurations are initialized with pre-trained full-precision network parameters, as is standard in [7]. The bit-width of weights and activations is denoted as I_W and I_A. From Tab. 1, we observe that one particular hyper-parameter configuration (η=1e-3, φ=1e-3 and τ=1e-7) is consistently outperforming the others on CIFAR-10 for both networks ResNet-20/56. For ImageNet, both AMSGrad configurations significantly outperformed all four ADAM optimizer configurations. This aligns with existing literature suggesting AMSGrad for the complex task of ImageNet [4, 8].

S2. Pruning Specific Training Parameters

Start ($\mathtt{t}_{\text{start}}$) and end ($\mathtt{t}_{\text{end}}$) of pruning, are training specific hyperparameters, which define the warm-up phase, the pruning phase, and the fine-tuning phase of PaBT. The total epochs are taken from [4], then the choice of pruning start and end points was done empirically, such that sufficient epochs are dedicated for the MBN to warm-up to a reasonable accuracy, followed by a long enough pruning stage that enables the *gradual* convergence down to a BNN. Finally, in the fine-tuning stage we use the remaining epochs to retrain until the accuracy is recovered. We found that extending the total number of epochs did not result in improved accuracy.

S3. Pruning as a Binarization Technique for Semantic Segmentation

Semantic segmentation is a crucial task which provides pixel-wise predictions in many application fields such as robotics and autonomous driving. Due to typically larger input image resolutions and additional layers in network architectures (bottleneck, Atrous Spatial Pyramid Pooling (ASPP) block and decoder layers), semantic segmentation surpasses the computational complexity of image classification. We show the scalability of PaBT to the task of semantic segmentation, where we adopt the DeepLab-based CNN architecture [1] with a ResNet-18 backbone. The last two residual blocks use a dilation rate of 2, while the ASPP blocks incorporate dilation rates {1, 8, 12, 18}. For all experiments, we set the input image resolution to

Table 1. Influence of the binary optimizer (BOP) training hyper-parameters, adaptivity rate φ and threshold τ, to train multi-bit networks in terms of Top-1.

Model/ Dataset	Optimizer	I_W/I_A	η	BOP Parameter φ	BOP Parameter τ	Top-1 [%]
ResNet-20 CIFAR-10	SGD (θ)	8/8	0.1	-	-	92.4
	ADAM ($\theta_{\mathrm{BN,SF}}$), BOP (θ_{W})	3/3	1e-2	1e-3	1e-6	89.63
					1e-7	89.17
					1e-8	88.82
				1e-4	1e-6	86.84
					1e-7	89.73
			1e-3	1e-3	1e-6	89.74
					1e-7	**90.00**
					1e-8	89.07
ResNet-56 CIFAR-10	SGD (θ)	8/8	0.1	-	-	93.89
	ADAM ($\theta_{\mathrm{BN,SF}}$), BOP (θ_{W})	1/1	1e-2	1e-3	1e-6	83.95
					1e-7	81.96
				1e-4	1e-8	86.78
			1e-3	1e-3	1e-6	87.40
					1e-7	**87.52**
				1e-4	1e-8	87.34
	ADAM ($\theta_{\mathrm{BN,SF}}$), BOP (θ_{W})	3/3	1e-2	1e-3	1e-6	89.91
					1e-7	89.63
					1e-8	89.07
				1e-4	1e-7	89.86
					1e-8	89.31
			1e-3	1e-3	1e-6	90.73
					1e-7	**91.74**
				1e-4	1e-7	90.47
ResNet-18 ImageNet	SGD (θ)	8/8	0.1	-	-	69.30
	ADAM ($\theta_{\mathrm{BN,SF}}$), BOP (θ_{W})	3/3	[2.5e-3, 5e-6]	[1e-4, 1e-6]	1e-7	59.78
					1e-8	58.01
			[2.5e-3, 5e-8]	[1e-4, 1e-8]	1e-7	58.63
					1e-8	60.00
	AMSGrad ($\theta_{\mathrm{BN,SF}}$), BOP (θ_{W})	3/3	[2.5e-3, 5e-8]	[1e-4, 1e-8]	1e-8	62.60
			[1e-3, 5e-8]			**62.86**

512 $\times$ 1024, where we quantize the ResNet-18 backbone as well as the decoder layers as they hold the majority of computational complexity. Tab. 2 presents the investigation of base-oriented (α and β) and operation-oriented (γ) scaling factors, different optimizer settings and PaBT-based quantization of MBNs, on the semantic segmentation dataset CityScapes [2] in terms of bit-widths and mIoU. PaBT shows its improvements when compared to experiments with network parameters trained using AMSGrad [8]. PaBT also outperforms equivalent 1×1 models which use the binary optimizer (BOP) [4] to train weights in a latent-free manner, and train batch norm (θ_{BN}) and scaling factors (θ_{SF}) with AMSGrad. PaBT is able to produce dominating solutions (mIoU) through pruning an over-parameterized MBN from M=N=3 down to M=N=1, resulting in an improvement of 3.57 p.p. compared to directly learning a DeepLab with 1-bit for weights and activations with BOP training for weights and AMSGrad optimizer for θ_{BN} and θ_{SF}.

Table 2. Influence of the scaling factors θ_{SF}, the used optimizer and operation level pruning in terms of number of bit-operations (bit-OPs) and mIoU for the semantic segmentation task on CityScapes [2].

Model/ Dataset	θ_{SF}	Optimizer (Parameter Scope)	Operation Pruning	Bit-Width I_W	Bit-Width I_A	mIoU [%]
DeepLab CityScapes	-	ADAM (θ)	✗	8	8	68.53
	α, β	ADAM (θ)	✗	1	1	50.95
	γ	AMSGrad (θ)	✗			50.21
	γ	AMSGrad ($\theta_{BN,SF}$), BOP (θ_W)	✗			51.10
	α, β	AMSGrad (θ)	✗	3	3	60.15
	γ	AMSGrad (θ)	✗			59.85
	γ	AMSGrad ($\theta_{BN,SF}$), BOP (θ_W)	✗			60.61
	γ	AMSGrad ($\theta_{BN,SF}$), BOP (θ_W)	✓	1	1	**54.67**

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[4] Koen Helwegen, James Widdicombe, Lukas Geiger, Zechun Liu, Kwang-Ting Cheng, and Roeland Nusselder. Latent weights do not exist: Rethinking binarized neural network optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019. 1, 2

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[6] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. *University of Toronto*, 2012. 1

[7] Xiaofan Lin, Cong Zhao, and Wei Pan. Towards Accurate Binary Convolutional Neural Network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1

[8] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 1