Supplementary Material: Structured Binary Neural Networks for Accurate Image Classification and Semantic Segmentation

Bohan Zhuang[1] Chunhua Shen[1]* Mingkui Tan[2] Lingqiao Liu[1] Ian Reid[1]
[1]Australian Centre for Robotic Vision, The University of Adelaide
[2]South China University of Technology

S1. More ablation study on ImageNet classification

In this section, we continue the Sec. 4.3 in the main paper to provide more comparative experiments. We define more methods for comparison as follows: **GBD v1:** We implement with the group-wise binary decomposition strategy, where each base consists of one block. It corresponds to the approach described in Eq. (5) and is illustrated in Fig. S1 (a). **GBD v2:** Similar to GBD v1, the only difference is that each group base has two blocks. It is illustrated in Fig. S1 (b) and is explained in Eq. (6). **GBD v3:** It is an extreme case where each base is a whole network, which can be treated as an ensemble of a set of binary networks. This case is shown in Fig. S1 (d).

S1.1. Group space exploration

We are interested in exploring the influence of different group-wise decomposition strategies. We present the results in Table S1. We observe that by learning the soft connections between each block results in the best performance on ResNet-18. And methods based on hard connections perform relatively worse. From the results, we can conclude that designing compact binary structure is essential for highly accurate classification. What's more, we expect to further boost the performance by integrating with the NAS approaches as discussed in Sec. S2.

S1.2. Effect of the number of bases

We further explore the influence of number of bases K to the final performance in Table S2. When the number is set to 1, it corresponds to directly binarize the original full-precision network and we observe apparent accuracy drop compared to its full-precision counterpart. With more bases employed, we can find the performance steadily increases. The reason can be attributed to the better fitting of the floating-point structure, which is a trade-off between accuracy and complexity. It can be expected that with enough bases, the network should has the capacity to approximate the full-precision network precisely. With the multi-branch group-wise design, we can achieve high accuracy while still significantly reducing the inference time and power consumption. Interestingly, each base can be implemented using small resource and the parallel structure is quite friendly to FPGA/ASIC.

S2. More discussions

Relation to ResNeXt [8]: The homogeneous multi-branch architecture design shares some spirit of ResNeXt and enjoys the advantage of introducing a "cardinality" dimension. However, our objectives are totally different. ResNeXt aims to increase the capacity while maintaining the complexity. To achieve this, it first divides the input channels into groups and perform efficient group convolutions implementation. Then all the group outputs are aggregated to approximate the original feature map. In contrast, we first divide the network into groups and directly replicate the floating-point structure for each branch while both weights and activations are binarized. In this way, we can reconstruct the full-precision structure via aggregating a set of low-precision transformations for complexity reduction in the energy-efficient hardware. Furthermore, our structured transformations are not restricted to only one block as in ResNeXt.

Group-Net has strong flexibility: The group-wise approximation approach can be efficiently integrated with Neural Architecture Search (NAS) frameworks [3,4,7,12,13] to explore the optimal architecture. Based on Group-Net, we can further add number of bases, filter numbers, connections among bases into the search space. The proposed approach can also be combined with knowledge distillation strategy as in [6,11]. The basic idea is to train a target low-precision network alongside another pretrained full-precision guidance network. An additional regularizer is added to minimize the difference between student's and teacher's intermediate feature representations for higher accuracy. In this way, we expect to further decrease the number of bases

[*] C. Shen is the corresponding author.

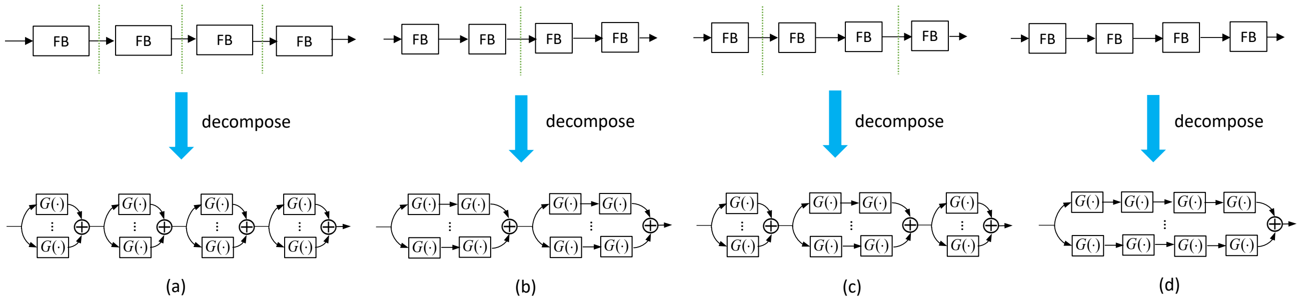

Figure S1: Illustration of several possible group-wise architectures. We assume the original full-precision network comprises four blocks. "FB" represents the floating-point block. $G(\cdot)$ is defined in Sec. 2.2.2 in the main paper, which represents a binary block. We omit the skip connections for convenience. (a): Each group comprises one block and we approximate each floating-point block with a set of binarized blocks. (b): Decompose the network into groups, where each group contains two blocks. Then we approximate each floating-point group using a set of binarized groups. (c): Each group contains different number of blocks. (d): An extreme case. We directly decompose the whole floating-point network into an ensemble of several binary networks.

Table S1: Comparisons between several group-wise decomposition strategies. Top-1 and Top-5 accuracy gap to the corresponding full-precision networks are also reported.

Model	Bases	Top-1 %	Top-5 %	Top-1 gap %	Top-5 gap %
ResNet-18 Full-precision	1	69.7	89.4	-	-
Group-Net	5	64.8	85.7	4.9	3.7
GBD v1	5	63.0	84.8	6.7	4.6
GBD v2	5	62.2	84.1	7.5	5.3
GBD v3	5	59.2	82.3	10.5	7.1

Table S2: Validation accuracy of Group-Net on ImageNet with different number of bases. All cases are based on the ResNet-18 network with binary weights and activations.

Model	Bases	Top-1 %	Top-5 %	Top-1 gap %	Top-5 gap %
Full-precision	1	69.7	89.4	-	-
Group-Net	1	56.4	79.5	13.3	9.9
Group-Net	3	62.5	84.2	7.2	5.2
Group-Net	5	64.8	85.7	4.9	3.7

while maintaining the performance.

S3. More ablation study on semantic segmentation

S3.1. Influence of dilation rates on full-precision baselines

In this section, we explore the effect of dilation rates in the last two blocks for full-precision baselines. We show the mIOU change in Figure. S2. For dilation rates (1, 1), it corresponds to the original FCN baseline [5] with no atrous convolution applied. For both FCN-32s and FCN-16s, we can observe that when using dilated convolution with $rate = 4$ and $rate = 8$ in the last two blocks respectively, we can get the best performance.

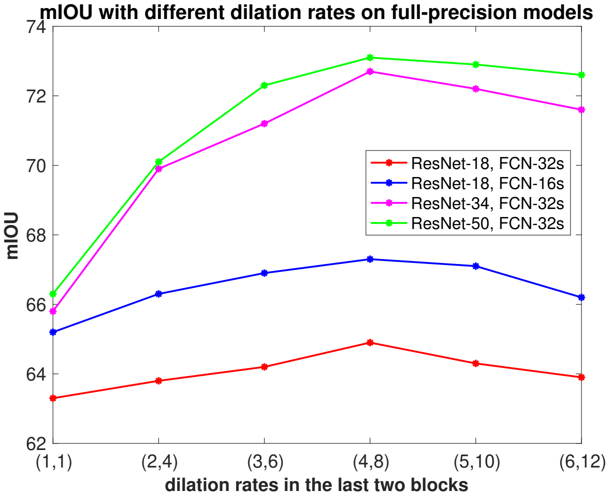

Figure S2: Illustration of the influence of different dilation rates in the last two blocks for the floating-point baseline models.

S3.2. Full-precision baselines with multiscale dilations

In Sec. 4.4 in the paper, we have shown that Group-Net with BPAC can accurately fit the full-precision model while saving considerable computational complexity. To explore the effect of multiscale dilations on full-precision models, we replace the last two blocks as the same structure of BPAC. Specifically, we use K homogeneous floating-point branches in the last two blocks while each branch is different in dilation rate. We set $K = 5$ here. Because of this modification, the FLOPs for full-precision ResNet-18, ResNet-34 and ResNet-50 increases by 2.79×, 3.14× and 3.13×, respectively. As shown in Table S3, the multiple dilations design improves the performance of full-precision

Table S3: Performance on PASCAL VOC 2012 validation set.

Model		mIOU
ResNet-18, FCN-32s	Full-precision (multi-dilations)	67.6
	Full-precision	64.9
	Group-Net + BPAC	63.8
ResNet-18, FCN-16s	Full-precision (multi-dilations)	70.1
	Full-precision	67.3
	Group-Net + BPAC	66.3
ResNet-34, FCN-32s	Full-precision (multi-dilations)	75.0
	Full-precision	72.7
	Group-Net + BPAC	71.2
ResNet-50, FCN-32s	Full-precision (multi-dilations)	75.5
	Full-precision	73.1
	Group-Net + BPAC	70.4

baselines but at a cost of huge computational complexity increase. In contrast, Group-Net + BPAC does not increase the computational complexity compared with using Group-Net only. This proves the flexibility of the proposed Group-Net which can effectively borrow task-specific properties to approximate the original floating-point structure. And this is one of the advantages for employing structured binary decomposition.

S4. Extending Group-Net to binary weights and low-precision activations

In the main paper and in Sec. S1 to Sec. S3, all the experiments are based on binary weights and binary activations. To make a tradeoff between accuracy and computational complexity, we can add more bases as discussed in Sec. S1.2. However, we can also increase the bit-width of activations for better accuracy according to actual demand. We conduct experiments on the ImageNet dataset and report the accuracy in Table S4, Table S5 and Table S6.

S4.1. Fixed-point Activation quantization

We apply the simple uniform activation quantization in the paper. As the output of the ReLU function is unbounded, the quantization after ReLU requires a high dynamic range. It will cause large quantization errors especially when the bit-precision is low. To alleviate this problem, similar to [2, 10], we use a clip function $h(y) = \mathrm{clip}(y, 0, \beta)$ to limit the range of activation to $[0, \beta]$, where β (not learned) is fixed during training. Then the truncated activation output $\mathbf{y}$ is uniformly quantized to K-bits ($K > 1$) and we still use STE to estimate the gradient:

$$\begin{aligned} \text{Forward} : \widetilde{\mathbf{y}} &= \mathrm{round}(\mathbf{y} \cdot \frac{2^K - 1}{\beta}) \cdot \frac{\beta}{2^K - 1}, \\ \text{Backward} : \frac{\partial \ell}{\partial \mathbf{y}} &= \frac{\partial \ell}{\partial \widetilde{\mathbf{y}}}. \end{aligned} \tag{1}$$

Since the weights are binary, the multiplication in convolution is replaced by fixed-point addition. One can simply replace the uniform quantizer with other non-uniform quantizers for more accurate quantization similar to [1, 9].

S4.2. Implementation details

For data preprocessing, it follows the same pipeline as BNNs. We also quantize the weights and activations of all convolutional layers except that the first layer and the last layer are full-precision. For training ResNet with fixed-point activations, the learning rate starts at 0.05 and is divided by 10 when it gets saturated. We use Nesterov momentum SGD for optimization. The mini-batch size and weight decay are set to 128 and 0.0001, respectively. The momentum ratio is 0.9. We directly learn from scratch since we empirically observe that fine-tuning does not bring further benefits to the performance. The convolution and element-wise operations are in the order: Conv → BN → ReLU → Quantize.

S4.3. Evaluation on ImageNet

For experiments in Table S4 and Table S5, we use 5 bases (*i.e.*, $K = 5$). From Table S4, we can observe that with binary weights and fixed-point activations, we can achieve highly accurate results. For example, by also referring to Table 2 in the main paper, we can find the Top-1 accuracy drop for Group-Net on ResNet-50 with tenary and binary activations are 1.5% and 6.5%, respectively. Furthermore, our approach still works well on plain network structures such as AlexNet in Table S5. We also provide the comparison with different number of bases in Table S6.

Table S4: Validation accuracy of different binary decomposition strategies on ImageNet with different choices of W and A. 'W' and 'A' refer to the weight and activation bitwidth, respectively.

Model	W	A	Top-1 %	Top-5 %	Top-1 gap %	Top-5 gap %
ResNet-18 Full-precision	32	32	69.7	89.4	-	-
Group-Net	1	2	69.6	89.0	0.1	0.4
Group-Net	1	32	70.4	89.8	-0.7	-0.4
GBD v1	1	4	69.2	88.5	0.5	0.9
GBD v2	1	4	68.3	87.9	1.4	1.5
GBD v3	1	4	64.5	85.0	5.2	4.4
LBD	1	4	60.1	82.2	9.6	7.2
ResNet-50 Full-precision	32	32	76.0	92.9	-	-
Group-Net	1	2	74.5	91.5	1.5	1.4
Group-Net	1	4	76.0	92.7	0.0	0.2

Table S5: Accuracy of AlexNet on ImageNet validation set. All cases use binary weights and 2-bit activations.

Model	Full-precision	LBD	GBD v1	Group-Net
Top-1 %	57.2	54.2	57.3	**57.8**
Top-5 %	80.4	77.6	80.1	**80.9**

References

[1] Z. Cai, X. He, J. Sun, and N. Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5918–5926, 2017. 3

Table S6: Validation accuracy of Group-Net on ImageNet with number of bases. All cases are based on the ResNet-18 network with binary weights and 4-bit activations.

Model	Bases	bitW	bitA	Top-1 %	Top-5 %	Top-1 gap %	Top-5 gap %
Full-precision	1	32	32	69.7	89.4	-	-
Group-Net	1	1	4	61.5	83.2	8.2	6.2
Group-Net	3	1	4	68.5	88.7	1.2	0.7
Group-Net	5	1	4	70.1	89.5	-0.4	-0.1

[2] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *J. Mach. Learn. Res.*, 18(1):6869–6898, 2017. 3

[3] C. Liu, B. Zoph, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy. Progressive neural architecture search. In *Proc. Eur. Conf. Comp. Vis.*, 2018. 1

[4] H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. 1

[5] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3431–3440, 2015. 2

[6] A. Mishra and D. Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *Proc. Int. Conf. Learn. Repren.*, 2018. 1

[7] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean. Efficient neural architecture search via parameter sharing. In *Proc. Int. Conf. Mach. Learn.*, 2018. 1

[8] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5987–5995, 2017. 1

[9] D. Zhang, J. Yang, D. Ye, and G. Hua. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In *Proc. Eur. Conf. Comp. Vis.*, 2018. 3

[10] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016. 3

[11] B. Zhuang, C. Shen, M. Tan, L. Liu, and I. Reid. Towards effective low-bitwidth convolutional neural networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. 1

[12] B. Zoph and Q. V. Le. Neural architecture search with reinforcement learning. In *Proc. Int. Conf. Learn. Repren.*, 2017. 1

[13] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. 1