Supplementary material for BATS: Binary ArchitecTure Search

Adrian Bulat[1[0000−0002−3185−4979]], Brais Martinez[1[0000−0001−7511−8941]], and Georgios Tzimiropoulos[1,2[0000−0002−1803−5338]]

[1] Samsung AI Center, Cambridge, UK
adrian@adrianbulat.com, brais.mart@gmail.com
[2] Queen Mary University of London, London, UK
g.tzimiropoulos@qmul.ac.uk

A Additional comparison with state-of-the-art on CIFAR10

To further showcase the improvements offered by our approach, herein we compare its performance against an additional set of state-of-the-art methods on the CIFAR-10 dataset. As the results from Table 1 show, our method significantly outperforms all previous ones across different architectures(VGG, ResNet, WRN) and quantization levels.

Method	**Acc.(%)**	Architecture	# bits (W/A)
BC [3]	90.1	VGG-small	1/32
TTQ [12]	91.1	ResNet-20	2/32
HWGQ [1]	92.5	VGG-small	1/2
LQ-Net [11]	93.4	VGG-small	1/2
CBCN [5]	91.6	ResNet-18	(1/1)×4
CBCN [5]	93.4	WRN40	(1/1)×4
BNN [4]	89.9	VGG-small	1/1
XNOR-Net [7]	89.8	VGG-small	1/1
CCNN [10]	92.3	VGG-small	1/1
CI-Net [9]	92.5	VGG-small	1/1
BATS (Ours)	**96.1**	**BATS**	1/1

Table 1: Comparison with state-of-the-art binarization/quantization methods on CIFAR-10 across various architecture. Notice that the discovered architecture by our approach significantly outperforms all previous reported results.

B Going back to real

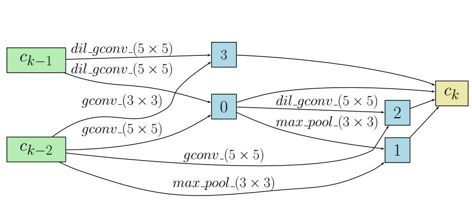
Herein we briefly evaluate the effectiveness of our novel search space and methodology for the case of real-valued networks. To do so, given the proposed search space and temperature regularization mechanism, we performed a network search on CIFAR-10 largely following the procedure described in Section 4, with the following changes: the learning rate for both search and evaluation is set to 0.1 and the optimizer to SGD with momentum 0.9. As Table 2 shows, our method generalizes well to the real-valued case, offering competitive results. This suggests that the operations tailored to binary networks can work well for their real-valued counterparts, too.

Table 2: Comparison on the CIFAR-10 dataset for the case of real-valued networks.

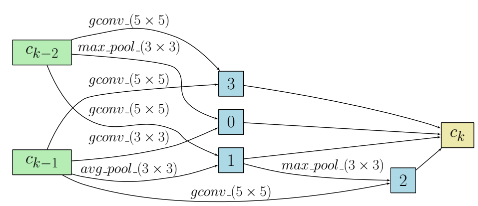
Architecture	**Test Err. (%)**	**Params (M)**
NASNet-A [13]	2.65	3.3
AmoebaNet-A [8]	3.34	3.2
DARTS (first order) [6]	3.00	3.3
DARTS (second order) [6]	2.76	3.3
P-DARTS [2]	2.50	3.4
BATS (Ours)	2.70	2.5
BATS (Ours)	2.40	3.5

C Discovered real-valued topologies

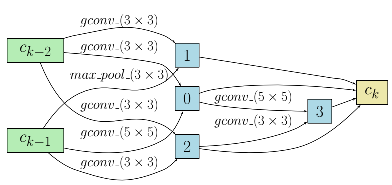
While the proposed search space and method is mainly geared towards binary networks, we also tested its generalizability on the real-valued domain. Fig. 1c and 1d depict an example of cells found by our approach when using real valued networks. Notice that as opposed to the binary ones (Fig. 1a and 1b), the real valued ones tend to be deeper and use operations with smaller convolutional kernels.



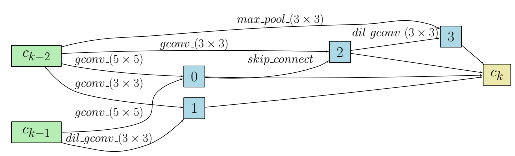
(a) Normal binary cell



(b) Reduction binary cell

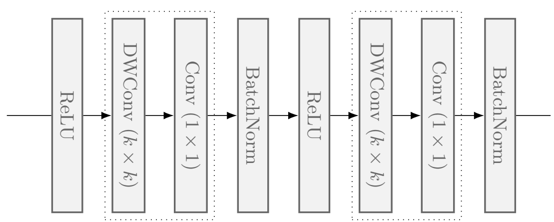


(c) Normal real-valued cell

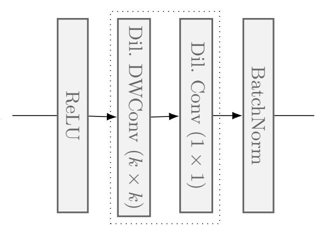


(d) Reduction real-valued cell

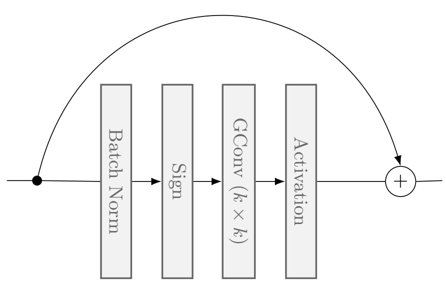
Fig. 1: Normal and reduction cells discovered by our proposed method using the introduced search space for the binary case (first row) and real-valued case (second row). Notice that the binary cells tend to be shallower and to contain convolutional operations with larger kernels (i.e. 5×5) when compared with the real-valued ones.



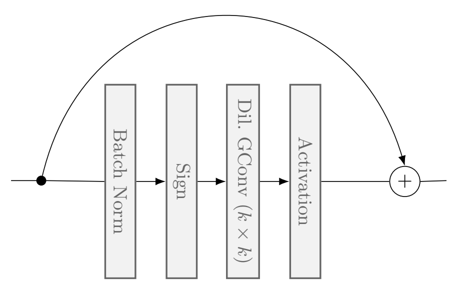
(a) DARTS depth-wise separable cell



(b) DARTS dilated convolution cell



(c) Proposed grouped convolution binary cell



(d) Proposed dilated grouped convolution binary cell

Fig. 2: Comparison between the convolutional operations used in the DARTS search space (2a and 2b) and the proposed ones (2c and 2d). $k \times k$ denotes the kernel size, $\bigoplus$ is the element-wise summation operation while each rectangle represents a given operation defined by the inner text.

References

1. Cai, Z., He, X., Sun, J., Vasconcelos, N.: Deep learning with low precision by half-wave gaussian quantization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5918–5926 (2017)
2. Chen, X., Xie, L., Wu, J., Tian, Q.: Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. IEEE International Conference on Computer Vision (2019)
3. Courbariaux, M., Bengio, Y., David, J.P.: Binaryconnect: Training deep neural networks with binary weights during propagations. In: Advances on Neural Information Processing Systems (2015)
4. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. arXiv preprint arXiv:1602.02830 (2016)
5. Liu, C., Ding, W., Xia, X., Zhang, B., Gu, J., Liu, J., Ji, R., Doermann, D.: Circulant binary convolutional networks: Enhancing the performance of 1-bit dcnns with circulant back propagation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2691–2699 (2019)
6. Liu, H., Simonyan, K., Yang, Y.: DARTS: Differentiable architecture search. International Conference on Learning Representations (2019)
7. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: XNOR-Net: ImageNet classification using binary convolutional neural networks. In: European Conference on Computer Vision. pp. 525–542 (2016)
8. Real, E., Aggarwal, A., Huang, Y., Le, Q.V.: Regularized evolution for image classifier architecture search. In: AAAI Conf. on Artificial Intelligence. vol. 33, pp. 4780–4789 (2019)
9. Wang, Z., Lu, J., Tao, C., Zhou, J., Tian, Q.: Learning channel-wise interactions for binary convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 568–577 (2019)
10. Xu, Z., Cheung, R.C.: Accurate and compact convolutional neural networks with trained binarization. arXiv preprint arXiv:1909.11366 (2019)
11. Zhang, D., Yang, J., Ye, D., Hua, G.: Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In: European Conference on Computer Vision. pp. 365–382 (2018)
12. Zhu, C., Han, S., Mao, H., Dally, W.J.: Trained ternary quantization. arXiv preprint arXiv:1612.01064 (2016)
13. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8697–8710 (2018)