A Training Configuration of Pilot Study

In our pilot study, we follow the typical training settings to train binary MLP-Mixer, ResNet-18 and the ResNet+MLP architectures. Details are shown in Table 1.

Table 1: Training settings of pilot study on the ImageNet1k benchmark.

config	value
optimizer	AdamW [4]
learning rate	0.001
weight decay	0.01
batch size	1024
learning rate schedule	cosine decay [3]
warmup iterations	6250
training iterations	125000 ($\approx$ 100 epochs)
label smooth	0.1
distillation	none
two-step training [5]	none

B Distillation Configuration

Knowledge distillation [1] is widely used in BNN training, with the real-valued model as a teacher and the 1-bit network as a student. For example, Real-to-Binary [5] employs a multilevel distillation loss to learn from the middle stage of a teacher model. ReActNet [2] simplifies distillation and only applies KL divergence in the last layer. In this work, we train BCDNets with both distillation and label-supervision losses. In training, we attach two fully-connected layers at last to learn from the teacher and real labels based on the KL divergence and cross-entropy loss respectively. Given results $\boldsymbol{y}_t$ from a teacher network and global-average-pooled features $\boldsymbol{y}_s$ from a student network, the overall objective is formulated as:

$$L(\boldsymbol{X}) = \frac{1}{2}L_{CE}(\boldsymbol{W}_{label}^T\boldsymbol{y}_s + \boldsymbol{b}_{label}, \boldsymbol{y}_{label}) + \frac{1}{2}KL(\boldsymbol{W}_{dist}^T\boldsymbol{y}_s + \boldsymbol{b}_{dist}, \boldsymbol{y}_t), \tag{1}$$

where $\{\boldsymbol{W}_{label}^T, \boldsymbol{b}_{label}\}$ and $\{\boldsymbol{W}_{dist}^T, \boldsymbol{b}_{dist}\}$ indicate weights and biases in the label-supervision head and distillation head. We re-parameterize two heads at inference. During testing, the linear transformation in both heads can be merged as:

$$\boldsymbol{W} = \boldsymbol{W}_{label} + \boldsymbol{W}_{dist}, \quad \boldsymbol{b} = \boldsymbol{b}_{label} + \boldsymbol{b}_{dist}, \tag{2}$$

$$\texttt{BCDNet}(\boldsymbol{X}) = \frac{1}{2}\boldsymbol{W}^T\boldsymbol{y}_s + \frac{1}{2}\boldsymbol{b}. \tag{3}$$

As such, zero operations and zero parameters increased but enjoying double model capacity for training in the last layer.

Table 2: Efficacy of distillation, where "†" indicates results cited from ReActNets [2]. We also report the result of ReActNet-A with our training settings.

Method	Distillation	2 heads	Top1	Top5
ReActNet-Baseline†	✗	–	61.1†	–
BCDNet-A	✗	–	69.31	88.41
ReActNet-A†	✓	–	69.4†	–
ReActNet-A	✓	✓	70.31	89.05
BCDNet-A	✓	✗	71.66	90.28
BCDNet-A	✓	✓	**71.76**	**90.32**

Table 3: Training settings on 5 fine-grained small datasets.

config	value
optimizer	AdamW [4]
learning rate	0.0005
weight decay	0.01
batch size	512
learning rate schedule	cosine decay [3]
training epochs	100
label smooth	0.1
distillation	none
two-step training [5]	none

In Table 2, we evaluate the efficacy of distillation in BCDNets. During training, we choose the real-valued ResNet50 as the distillation teacher. First, we train a BCDNet-A without distillation as the baseline. Second, we report the result of ReActNet-A training in our settings with two distillation heads for comparison. Finally, we explore the efficacy of our re-parameterizable distillation heads, where "2 heads" indicates training with two heads for label supervision and distillation respectively. For comparison, we also train the models with a single head for both label and teacher supervision. As in previous works, the distillation is necessary for training binary networks, which improves 2+% accuracy. We retrain ReActNet-A in our training setting with two distillation heads. BCDNet-A exceeds ReActNet-A 1.3% top 1 accuracy with replacements of binary contextual MLPs. When it comes to the independent distillation and label-supervision heads, performance slightly improves 0.1% accuracy in BCDNet-A, while reparameterization guarantees no additional cost at inference.

C Training Configuration for Fine-Grained Datasets

We train different binary neural networks on 5 fine-grained small datasets including CUB-200-2011, Oxford-flowers102, Aircraft, Stanford-cars, Stanford-dogs. The detailed training setting is summarized in Table 3.

References

1. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)
2. Liu, Z., Shen, Z., Savvides, M., Cheng, K.T.: Reactnet: Towards precise binary neural network with generalized activation functions. In: European Conference on Computer Vision (ECCV) (2020)
3. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
4. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
5. Martinez, B., Yang, J., Bulat, A., Tzimiropoulos, G.: Training binary neural networks with real-to-binary convolutions. arXiv preprint arXiv:2003.11535 (2020)