

Multi-bit Adaptive Distillation for Binary Neural Networks–Supplementary File

Ying Nie
ying.nie@huawei.com

Noah's Ark Lab,
Huawei Technologies

Kai Han
kai.han@huawei.com

Yunhe Wang
yunhe.wang@huawei.com

1 Ablation Studies

Multi-teachers v.s. Single-teacher. We firstly compare the performance of different single teacher distillation and multiple teachers distillation using ResNet-20 on CIFAR-10 dataset and ResNet-18 on ImageNet dataset. From Table 2, we can see that the accuracy of multi-teachers distillation surpasses all single teacher distillation for both networks. Compared with the best single-teacher distillation, there are 0.58% and 0.9% improvement for ResNet-20 and ResNet-18 respectively. For the lightweight ReActNet-A, MAD also has a 0.8% improvement over the best single-teacher distillation. Besides, the standard deviation of our method on CIFAR-10 is significantly lower than that of other single-bit teachers except 2-bit teacher.

Table 1: Comprehensive ablation study results using ResNet-20 on CIFAR-10.

Method	FULL	w/o AKA	w/o CTL	w/o Middle	w/o Final
Top1(%)	85.91 \pm 0.12	85.48 \pm 0.10	84.92 \pm 0.33	85.47 \pm 0.25	84.52 \pm 0.09

Adaptive Knowledge Adjusting. We compare the performance of our method with and without adaptive knowledge adjusting scheme in Table 1, which are represented by "FULL" and "w/o AKA", respectively. For "w/o AKA", we fix the coefficients of four teachers as 1 all the training process. The final classification accuracy is 85.48%, 0.43% lower than the adaptive method.

Table 2: Accuracy of single-teacher distillation and multi-teachers distillation.

Teacher	ResNet-20 CIFAR-10	ResNet-18 ImageNet	ReActNet-A ImageNet
FP	84.94 \pm 0.17	57.0	69.4
8-bit	85.07 \pm 0.21	56.9	69.2
4-bit	84.95 \pm 0.21	56.8	68.9
2-bit	85.33 \pm 0.08	57.0	69.0
MAD	85.91 \pm 0.12	57.9	70.2

Convolutional Transforming Layer. Table 1 shows the result achieved by the method without 1×1 convolutional transforming layer which is denoted as "w/o CTL". For "w/o CTL", the student network intermediate layers' features is directly distilled by multiple teacher networks without transformation, and the classification accuracy is reduced by 0.99%. Besides, the training gets slightly unstable without the transforming layer.

Distillation by Different Layers. We also compare the performance distilled by different layers as shown in Table 1. "w/o Middle" means only using the knowledge of the final classification layer and "w/o Final" means only using the knowledge of intermediate layers, which achieve 85.47% and 84.52% classification accuracy, respectively. Comparatively, distillation without the classification layer will leads to a higher precision degradation. The knowledge of intermediate layers can help stabilize the distillation process and improve the final precision of student a little bit.