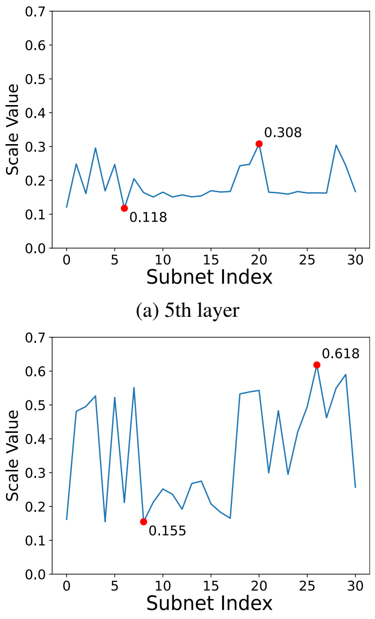
QuantNAS: Quantization-aware Neural Architecture Search For Efficient Deployment On Mobile Device

Supplementary Material

Due to the space limitation of the main paper, we will include additional analysis and experimental results in this supplementary file. In Sec. 7, we provide the experimental analysis of the optimal scale in different subnets. Sec. 8 is the searching results on Cifar10 and Cifar100 datasets.

7. Optimal scale in different subnets

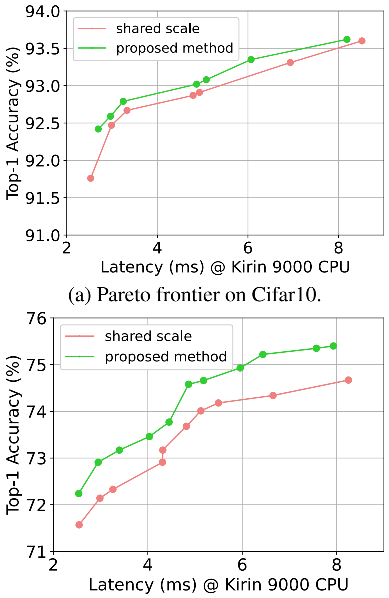
As mentioned in Sec. 3.2 in the main paper, the folded weights of each subnet are different after BN calibration. To obtain the optimal scale of each subnet, we randomly select 30 subnets from our search space and train them from scratch. Figure 12, demonstrates the optimal scale of 5th layer and 48th layer in the selected subnets. As can be seen, the maximum scale is 3 times of the minimum scale in 5th layer. The ratio increase to 4 times in the 48th layer. This comparison demonstrates that the optimal scale of each subnet has large difference. Training the supernet with a shared scale which is not the best for each subnet will result in non-optimal searching results. Therefore, a shared scale is not suitable for QuantNAS.



(a) 5th layer

(b) 48th layer

Figure 12. Optimal scale value for different subnets on the same layer.



(a) Pareto frontier on Cifar10.

(b) Pareto frontier on Cifar100.

Figure 13. Pareto frontier of training with shared scale and the proposed scale predictor.

8. Searching results on Cifar10 and Cifar100

To demonstrate the effectiveness of the proposed scale predictor, we verify the Pareto frontier of training the supernet with scale predictor and the one training with shared scale on ImageNet sub-100, Cifar10 and Cifar100 datasets. Results on ImageNet sub-100 have been presented in Sec. 5.3.2 in the main paper. Results on Cifar10 and Cifar100 are illustrated in Figure 13. As can be seen, the performance of the searched architectures from the proposed QuantNAS is superior to the counterpart of training with shared scale. This results show that the scale predictor has generalizability on different datasets.