USDN: A Unified Sample-wise Dynamic Network with Mixed-Precision and Early-Exit

Supplementary Materials

Appendix A

Training scheme for USDN

We pre-trained the full-precision supernet and then fine-tuned it through quantization-aware training (QAT). The network was trained for 300 epochs using an SGD optimizer with a weight decay of 1e-4. The learning rate started at 0.1 and decreased using the cosine annealing learning rate scheduler. When retraining the searched network, we trained a full-precision teacher network with CE loss. Subsequently, we trained a USDN using Eq. (10). During this phase, the weight decay was changed to 5e-4. The optimal temperature T was found to be 2, while α was set to 0.3. We applied an identical training process for both datasets except for the batch size. For ImageNet, the batch size was set to 1024, and for CIFAR, the batch size was set to 64.

Appendix B

Training scheme for quantized EEnets

We trained a 4-bit quantized MSDnet [2] with a model configuration of block=5 and step=4. When training the MSDnet, we loaded the full-precision pretrained weights from the GitHub repository (https://github.com/kalviny/MSDNet-PyTorch) and subsequently fine-tuned them using QAT. Initially, we followed the training scheme from [2], but it exhibited accuracy degradation. Consequently, we applied the same training scheme that we used for training our USDN, as elaborated in the preceding section.

In the case of SDN [3], we trained from scratch in full precision and then subsequently fine-tuned it. Once again, we applied our training scheme instead of the one presented in [3] because ours demonstrated better accuracy.

Appendix C

Exit Rate and computation cost of ICs

In this section, we present the exit rates and inference costs associated with each classifier, as reported in Table 1 of the manuscript. The numerical results are provided in Tab. I. Exit rates and inference costs vary based on the exit threshold. In the Tab. I, we display the results for the case where the threshold is set to 0.7.

Appendix D

Effect of Level-wise Knowledge Distillation Training

In this section, we examine how the proposed knowledge-distillation (KD) training improves the accuracy-BOPs trade-offs of the quantized early-exit network (EEnet). First, we conduct an analysis of how KD affects the confidence calibration [1] of each internal classifier (IC). The concept of confidence calibration has been introduced to ensure that a model's probability reflects its actual accuracy. This is highly relevant to EEnet's performance because the early-exit policy is based on the confidence value of the IC. If the model is underconfident, it cannot promptly exit easy-to-classify samples, which leads to an increase in computational costs. Interestingly, our findings reveal that the quantized EEnet tends to be underconfident in its predictions. To quantify the reliability of the confidence value, we calculate the expected calibration error (ECE) [4] and maximum calibration error (MCE) [4] of each IC. In Tab. II, we list the variance of error rate, ECE, and MCE when KD training is applied to the EE+4MP. It can be observed that both ECE and MCE have improved across all classifiers.

Appendix E

Searched Configuration

In this section, we provide our search space and searched configuration of each model. For ResNet18, the possible candidate bit-width of each residual block is set to $\{(W,A)\} = \{(2,2), (3,3), (4,4), (5,5)\}$. We present the search result with Fig. 1, Fig. 2, and Fig. 3.

TABLE I. Early-exit rate and inference cost consumed at each classifier. USDNs are searched and evaluated on the *ImageNet* dataset.

Model	Exit0		Exit1		Exit2	
	Exit Rate (%)	Cost (G)	Exit Rate (%)	Cost (G)	Exit Rate (%)	Cost (G)
USDN-tr03(th=0.7)	5.05	0.459	28.77	4.041	66.18	12.059
USDN-tr02(th=0.7)	6.19	0.634	24.58	3.736	69.23	14.215
USDN-tr01(th=0.7)	6.47	0.662	30.61	5.360	62.93	16.182

TABLE II. Comparison between the proposed knowledge distillation loss and cross-entropy loss. The experiment is conducted on USDN-tr01 EE+4MP (ResNet18) on the *ImageNet* dataset.

	Exit0			Exit1			Exit2		
	Top.1 Err ↓	ECE↓	MCE↓	Top.1 Err ↓	ECE↓	MCE↓	Top.1 Err ↓	ECE↓	MCE↓
NonKD	0.659	0.110	0.233	0.376	0.102	0.153	0.298	0.071	0.113
KD	0.666	0.100	0.177	0.406	0.096	0.149	0.292	0.020	0.065
Variance (%)	+0.95	**-9.03**	**-24.03**	+8.15	**-6.06**	**-2.68**	**-1.83**	**-71.52**	**-42.72**

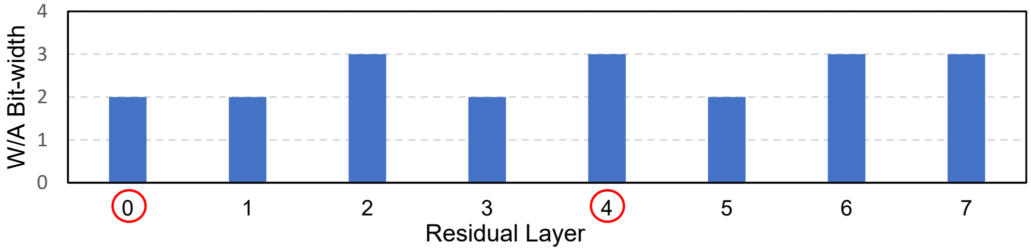

Fig. 1. Visualization of layerwise bit-width of EE+2.5MP (tr03) for ImageNet.

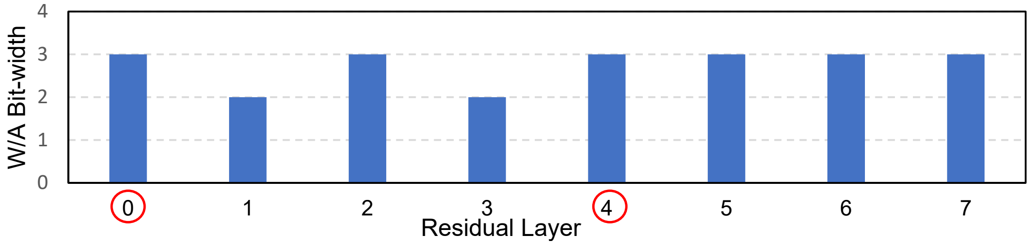

Fig. 2. Visualization of layerwise bit-width of EE+3MP (tr02) for ImageNet.

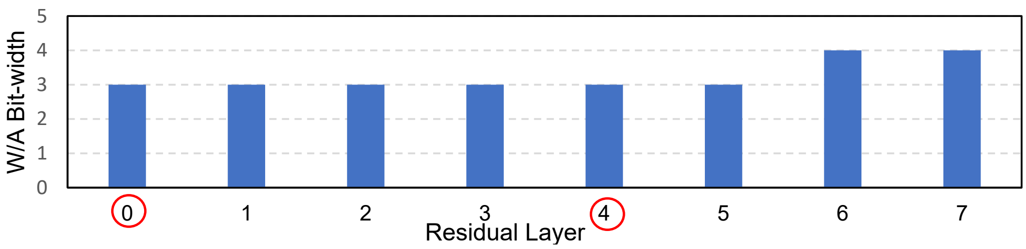

Fig. 3. Visualization of layerwise bit-width of EE+3MP (tr01) for ImageNet.

References

[1] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 2

[2] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification. In *ICLR*, 2018. 1

[3] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. Shallow-deep networks: Understanding and mitigating network overthinking. In *International Conference on Machine Learning (ICLR)*, pages 3301–3310. PMLR, 2019. 2

[4] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015. 2