NTIRE 2024 Efficient SR Challenge Factsheet -AdvancedSR-17

Yixuan Liu
Advanced Micro Devices, Inc
yixuanl@amd.com

Wensong Chan
Advanced Micro Devices, Inc
wensong.chan@amd.com

1. Team details

- Team name: AdvancedSR
- Team Leader's name: Yixuan Liu
- Team Leader's email: yixuanl@amd.com
- Rest of the team members: Wensong Chan, Dehua Tang
- Affiliation: Advanced Micro Devices, Inc.
- Affiliation of the team and/or team members with NTIRE 2024 sponsors (check the workshop website)
- User names and entries on the NTIRE 2024 Codalab competitions (development/validation and testing phases)
  lyx95lynn
- Best scoring entries of the team during development/validation phase
  test:27.021 valid.26.909
- Link to the codes/executables of the solution(s)
  https://github.com/lyxlynn/NTIRE2024_ESR

2. Method details

2.1. General method description.

The AdvancedSR team proposes a kernel-pruning based Residual Local Feature Network [2] for efficient SR. The overall network architecture is illustrated in Figure.1. It is a lightweight network consisting of a series of RLFB blocks, similar to RLFN. However, we prune the second convolution in pruned-RLFB block based on sensitivity analysis. Additionally, the pixelshuffle block is used for image restoration.

2.2. Pruning strategy.

Traditional channel-wise pruning methods struggle to effectively prune efficient super-resolution models. In comparison, we found that kernel pruning is more runtime-friendly and easier to recover the performance. When fine-tuning a subnet by directly removing weights, there's a risk of corrupting the original model weight. Therefore, we chose to use the center weight of the 3x3 convolutional kernel as the initialized weight for the pruned 1x1 convolutional layer, facilitating quicker convergence .

We performs the pruning steps based on NTIRE2024 official baseline model [4]. The pruning process consists of two stages. In the first stage, we apply progressive kernel-pruning on the re-parameterized model inspired by UPDP [5]. And in the second stage, we apply bias prune in our model to accelerate the runtime.

Stage 1. Kernel Pruning. We conducted sensitivity analysis on all the conv3x3 of RLFN Network. Based on the proportion of responses of the center weights in Conv3, we replaced the conv3 with conv1 progressively. The experiments show that replacing Conv3 with Conv1s scarcely degrades the performance but can enhance runtime efficiency. Additionally, We compare the efficiency of channel pruning with kernel pruning, detailed in Table 1. Despite channel pruning boasting fewer FLOPs, its runtime performance falls short of kernel pruning.

Stage 2.Bias Pruning. After kernel pruning, we obtain a subnet with a modified RLFB network. We preserve the biases of the first conv and ESA module , then retrain the subnet by removing the remaining biases thus further enhancing runtime efficiency.

2.3. Implementions details of AdvancedSR

The model is trained on LSDIR dataset [3] and we use RLFN(the official baseline model of NTIRE2024) as our basemodel. The training HR patch size is set as 256 x 256 with data augmentation such as rotation and horizontal flip in order to enhance the comprehensive ability of the model. We set the batch size as 64 in training process with totally

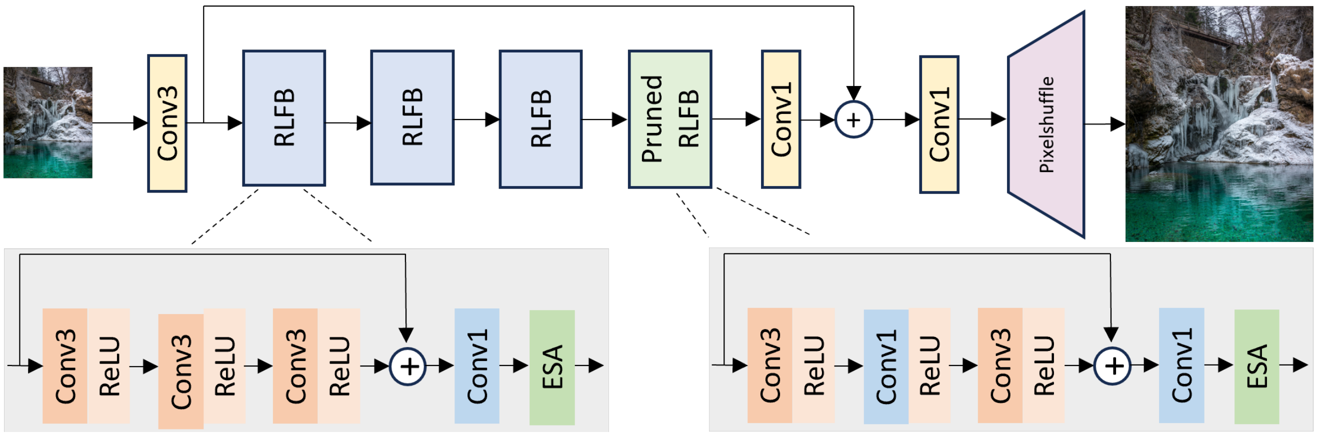

Figure 1. The overall architecture of AdvancedSR.

Table 1. Quantitative comparison with state-of-the-art RLFN-based methods on DIV2K_LSDIR validation runtime. CP indicates channel pruning, and KP means kernel pruning.

Model	Runtime[ms]	FLOPs[G]	Params[M]	Mem[MB]
MegSR [6]	5.51	14.90	0.243	495.91
RLFN	6.10	19.67	0.317	468.77
RLFN+KP (ours)	5.03	16.29	0.263	468.62
RLFN+CP+KP	5.24	14.33	0.235	468.52

500 epochs. The model is trained by minimizing L2 loss with Adam optimizer. The initail learning rate is set to 2e-5 and the learning rate is decayed by half at 100 epochs.

We apply our kernel pruning technique upon the base RLFN [2] model, denoted as RLFN+KP. Quantitative comparison with other RLFN-based methods is shown in Table 1. Runtime and memory cost is profiled on a single NVIDIA A100 GPU after a warmup [1] of 50 iterations.

Compared with the vanilla RLFN, our solution reduces the runtime by 14.1% (from 6.10ms to 5.24ms). We also compare with an RLFN pruned both channel- and kernel-wise (RLFN+CP+KP), and interestingly observe a moderate increase in runtime with regard to RLFN+KP, despite obvious reduction on FLOPs. This suggests that kernel pruning is more runtime friendly than channel pruning. Compared with the previous best preforming model MegSR [6], our simple solution RLFN+KP still shows a very competitive performance, surpassing it by saving 8.7% runtime (5.51 ms *vs*. 5.03 ms).

References

[1] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 2

[2] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–776, 2022. 1, 2

[3] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023. 1

[4] Yawei Li, Yulun Zhang, Radu Timofte, Luc Van Gool, Lei Yu, Youwei Li, Xinpeng Li, Ting Jiang, Qi Wu, Mingyan Han, et al. Ntire 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1959, 2023. 1

[5] Ji Liu, Dehua Tang, Yuanxian Huang, Li Zhang, Xiaocheng Zeng, Dong Li, Mingjie Lu, Jinzhang Peng, Yu Wang, Fan Jiang, et al. Updp: A unified progressive depth pruner for cnn and vision transformer. *AAAI*, 2024. 1

[6] Lei Yu, Xinpeng Li, Youwei Li, Ting Jiang, Qi Wu, Haoqiang Fan, and Shuaicheng Liu. Dipnet: Efficiency distillation and iterative pruning for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1701, 2023. 2