Supplementary Material
LowFormer: Hardware Efficient Design for Convolutional Transformer Backbones

Moritz Nottebaum[1]
nottebaum.moritz@spes.uniud.it

Matteo Dunnhofer[1]
matteo.dunnhofer@uniud.it

Christian Micheloni[1]
christian.micheloni@uniud.it

[1]University of Udine, Italy

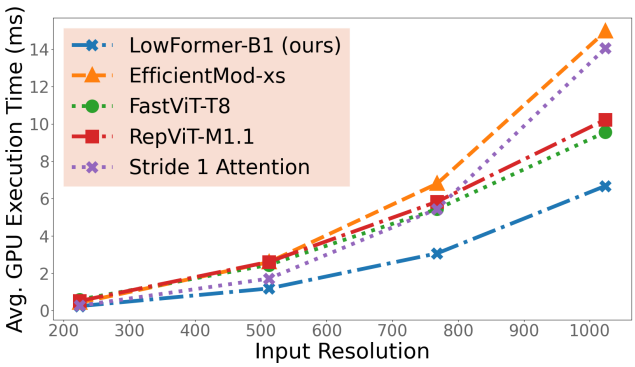

Figure 1. Average execution time for different models, depending on resolutions. The execution time is measured with batch size 200 and 5 warm-up iterations followed by 100 iterations from which the median is taken.

1. Latency Evaluation of fused vs unfused MBConv

In Figure 2 we measured relative latency of the fused and unfused MBConv [2]. The fused MBConv retains a lower latency for many configurations of channel dimension and operating resolution (value below 1 in Figure 2), even though its higher amount of MACs (value over 1 in Figure 2). Mainly for channel dimensions higher than 256, latency increases compared to the unfused MBConv. Some entries are missing due to OOM (out of memory) errors.

2. Resolution Scaling

In Figure 1 we evaluated execution time of various models with different input resolutions. We used a batch size of 200, the same as for throughput calculation. Even though LowFormer-B1 has a higher or similar top-1 accuracy, it retains a lower average execution time than EfficientMod-xs,

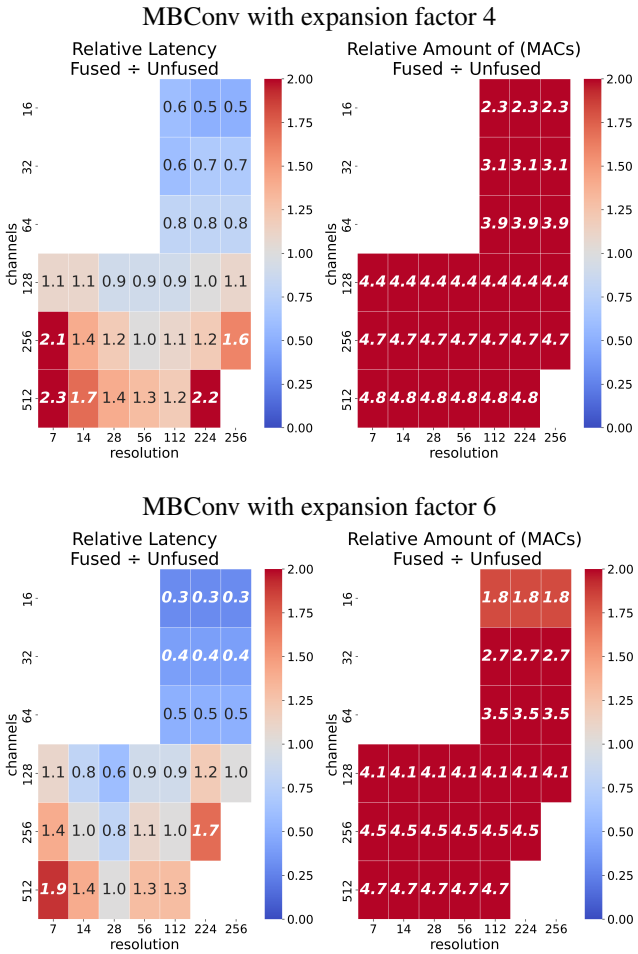

Figure 2. Latency comparison of fused and unfused MBConv. The missing fields resulted in out-of-memory errors.

[1] FastViT-T8 [3], RepViT-M1.1 [4] and the LowFormer-

B1 version without downsampling in the attention block, independent of the input resolution.

References

[1] Xu Ma, Xiyang Dai, Jianwei Yang, Bin Xiao, Yinpeng Chen, Yun Fu, and Lu Yuan. Efficient modulation for vision networks. *arXiv preprint arXiv:2403.19963*, 2024. 1

[2] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1

[3] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5785–5795, 2023. 1

[4] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15909–15920, June 2024. 1