LPViT: Low-Power Semi-structured Pruning for Vision Transformers —Supplementary Material

Kaixin Xu[1,2]*, Zhe Wang[1,2]*, Chunyun Chen[2], Xue Geng[1], Jie Lin[1], Xulei Yang[1], Min Wu[1], Xiaoli Li[1], and Weisi Lin[2]

[1] Institute for Infocomm Research (I^{2}R), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, 138632, Singapore
{xuk,wang_zhe,geng_xue,yang_xulei,wumin,xlli}@i2r.a-star.edu.sg
jie.dellinger@gmail.com
[2] College of Computing and Data Science (CCDS), Nanyang Technological University (NTU), Singapore
chunyun001@e.ntu.edu.sg,wslin@ntu.edu.sg

1 Experimental Details

Disccusions of parameters λ and β. We use following rule to adaptively determine β: $\beta^* = \sum_{l=1}^{L} \max_i \frac{\partial \delta_i}{\partial \alpha_i} / \sum_{l=1}^{L} \max_i \frac{\partial \mathcal{L}_{power}}{\partial \alpha_i}$. For λ, after fixing β, we adopt binary search to efficiently find the value of λ that results in targeted model FLOPs. The binary search requires only log time complexity. As a result, β controls the trade-off between distortion and power, λ controls global FLOPs. We further conduct an ablation study when setting β to different values shown in Tab. 1. As it shows, the adopted β selection strategies can find the best configuration in most cases, hence we adopt this strategy in other experiments as well.

Table 1: Ablation study of β selections on Swin-Tiny.

β	10^{-9}	8×10^{-7} (Reported)	10^{-6}	10^{-5}	10^{-4}
FLOPs (%)	71.96	**71.34**	71.42	71.63	71.73
Top-1	71.88	**79.24**	70.74	68.52	71.55

2 Qualitative Results of Segmentation Task.

A visualization of the qualitative performance of LPViT pruned segmentation model is also demontrated in Fig. 2, showing that the pruned model retains great visual quality even remaining only 50% FLOPs.

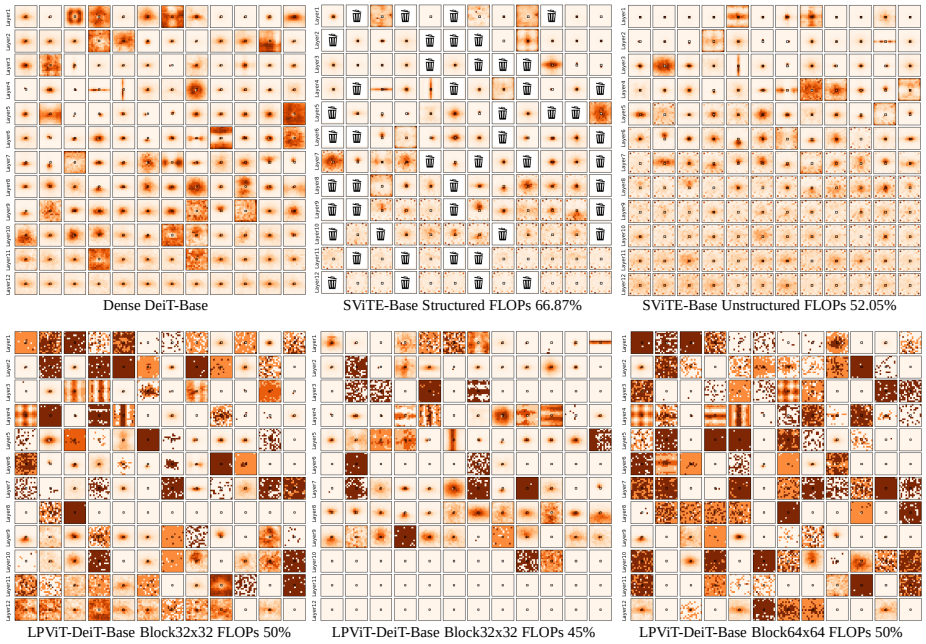

Fig. 1: Attention probabilities for DeiT-Base, SViTE-Base Structured/Unstructured pruning models as well as our LSP-DeiT-Base model with 12 layers (rows) and 12 heads (columns) using visualization tools provided in **[?]**.

Fig. 2: Qualitative result of segmentation masks on Cityscapes *val* set. First row shows ground truth mask and second row shows the predicted masks from pruned model.

3 Power as Constraint

To discuss the necessity of incoporating power penalty in the objective given the existing FLOPs constraints, we conduct an additional comparison with Power penalty-only approach in Tab. 2

Table 2: Hardware performances with only Power penalty.

Method	Sparsity = 28%		Sparsity = 51%		Sparsity = 75%	
	Speedup	Energy (W)	Speedup	Energy (W)	Speedup	Energy (W)
Power penalty only	2.4×	129(77.2%)	1.6×	152(91.1%)	1.23×	156(93.4%)
LPViT (Power+FLOPs)	3.87×	119(71.3%)	1.77×	140(83.8%)	1.29×	152(91%)

Tab. 2 shows that with only power penalty, energy is effectively reduced but the speedup is dropped. As a result, we use both power penalty and FLOPs constraint to obtain high energy savings and speedups simultaneously.