ELSA: Exploiting Layer-wise N:M Sparsity for Vision Transformer Acceleration

Supplementary Material

1. Appendix

1.1. Visualization of Search Results

In our study, we visually compare the sparsity levels identified by various algorithms for the Swin-Transformer, as depicted in Fig. 1. Observations from Fig. 1(b) and 1(c) reveal hierarchical trends in the N:M sparsity configuration for compression targets (*i.e.*, weight matrices of linear layers) identified by DominoSearch [5] and ER [4]. Specifically, these trends show lower sparsity (applying a denser choice, *e.g.*, 4:4 sparsity) in the layers of the initial blocks, with higher sparsity adopted in the deeper blocks. This pattern aligns with the pyramid-like structure of the Swin-Transformer, where the size of the linear layers increases progressively deeper into the model.

In contrast, the visualization of sparse configuration identified by our proposed ELSA framework, shown in Fig. 1(a), unveils a distinctive pattern. Here, higher sparsity (*e.g.,* 1:4)

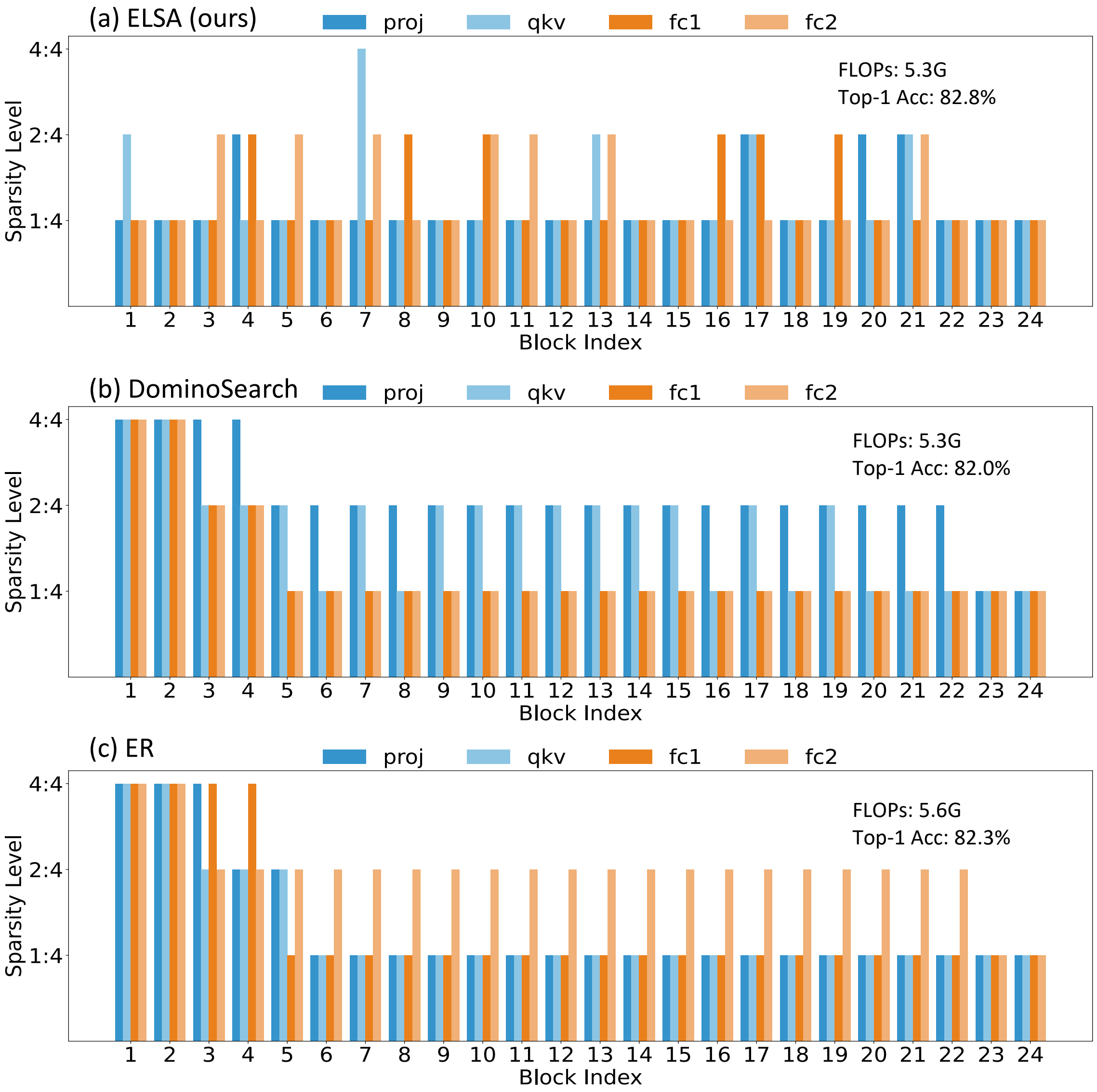

Figure 1. Visualization of sparsity configuration identified by different algorithms for the Swin-B model.

is applied to the initial blocks, while lower sparsity is selectively employed for the larger-sized linear layers in the model's middle and deeper sections. This strategic application of sparsity by the ELSA framework leads to a notable 0.8% improvement in Top-1 accuracy compared to DominoSearch, underscoring the critical importance and benefits of meticulously selecting sparsity levels for each compression target. Furthermore, it showcases the effectiveness of our ELSA framework in navigating these decisions.

1.2. Results on CNNs

Table 1. Experimental results of the proposed ELSA methodology on ConvNext-S. Accuracy denotes the Top-1 accuracy measure on the ImageNet-1K validation set.

Model	FLOPs	Accuracy
ConvNext-S	8.7G	82.8%
ELSA-ConvNext-S-2:4	4.3G (1.00×)	82.3%
ELSA-ConvNext-S-N:4	3.9G (1.11×)	82.2%
ELSA-ConvNext-S-N:4	3.5G (1.25×)	82.0%
ELSA-ConvNext-S-N:4	3.1G (1.41×)	81.6%

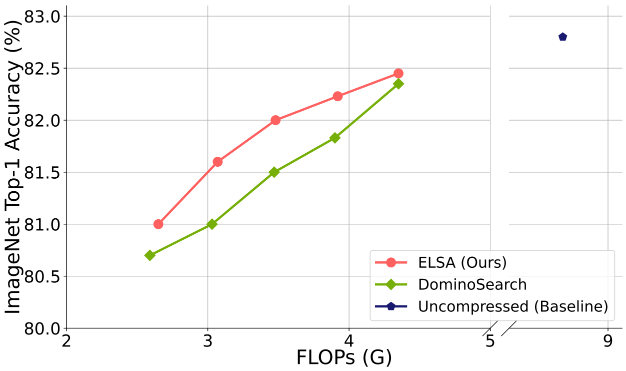

Figure 2. Comparisons of ELSA with DominoSearch on ConvNext-S

In this expanded investigation, we have applied the ELSA framework to the ConvNext-S model of convolution neural networks (CNNs) to explore its integration. Our goal is to demonstrate the comprehensive utility and effectiveness of the ELSA framework, showcasing its compatibility not only with vision transformers (ViTs) but also with a broader range of neural network architectures, including CNNs.

To validate this premise, we have conducted experiments on ConvNext-S, a notable advancement in CNN architecture. The experimental results in Table 1 confirm that the ELSA framework effectively adapts to the ConvNext-S model, significantly enhancing its computational efficiency while maintaining accuracy. This result is particularly significant, highlighting the versatility of the ELSA framework in accommodating various neural network paradigms, despite the operational distinctions between CNNs and ViTs. Furthermore, the N:M sparse networks searched by ELSA sit on the Pareto frontier, outperforming DominoSearch, as depicted in Fig. 2.

1.3. Ablation Study and Analysis

In this section, we present an ablation study analyzing the impact of our proposed supernet construction and sampling strategy on the resulting supernet quality. We conduct the experiment on the DeiT-S backbone. We visualize the result using the Pareto frontier analysis as shown in Fig. 3.

Effectiveness of Two-step Sampling Strategy In this ablation experiment, we adopt the vanilla sampling strategy [1], in which each sparse configuration is sampled uniformly to train a baseline supernet. The significant benefits of our two-step sampling strategy are illustrated in Fig. 3. Here, we can see that the proposed two-step sampling strategy can yield sparse subnetworks with superior performance at different levels of computation cost. Notably, at around the 1.6G FLOP mark, we can observe a nearly 1% improvement in Top-1 accuracy.

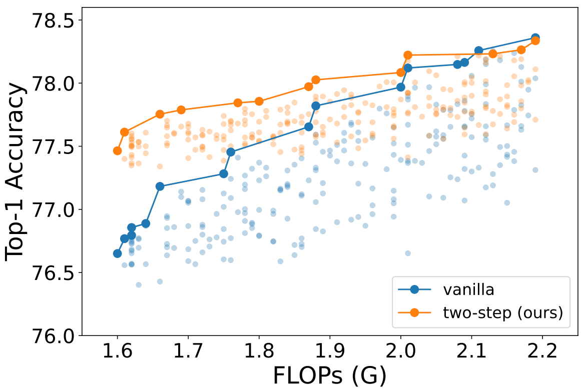

Figure 3. Pareto frontier of subnets randomly sampled from DeiT-S supernet trained with different sampling strategy

Table 2. Comparison of different estimators (supernet versus dense model)

Model	FLOPs	Estimator	Accuracy
ELSA-DeiT-B-N:4	7.0G	Dense model	81.0%
ELSA-DeiT-B-N:4	7.0G	Supernet	81.5%

Effectiveness of Supernet in Guiding Searching To demonstrate the benefit of using supernet, we run the evolutionary search and evaluate sparse configurations using lightweight metrics, i.e., the accuracy of the sparse NN, before fine-tuning. As in Table 2 below, the more accurate configuration ranking offered by the supernet can help identify a better sparse configuration with 0.5% higher accuracy.

Quality of Trained Supernet We aim to construct a high-quality sparse supernet, capable of empowering sparse networks within the design space needed to achieve a performance level close to fine-tuned networks. To evaluate our proposed sparse supernet, we employ the training paradigm for N:M sparse networks as ASP [3], refining the performance of sparse subnets through fine-tuning, starting with pretrained models. Each sparse subnet is meticulously fine-tuned for 150 epochs, maintaining consistent settings for knowledge distillation throughout the process. The comparison results are presented in Table 3. We note that the sparse subnets derived from our supernet demonstrate only a marginal decline in accuracy, ranging from 0.1% to 0.2% across various ViT backbones. These findings emphasize the effectiveness of our supernet training methodology. Through a single training cycle, a broad spectrum of sparse networks can be efficiently generated and inherited from our supernet, substantially reducing the fine-tuning costs.

Table 3. Comparison of 2:4 sparse subnets with inherited weights and 2:4 finetuned from pretrained weights (150 epochs)

Model	FLOPs	Inherited from ELSA	Fine-tuned from pretrained
DeiT-S	2.5G	79.1%	79.2%
DeiT-B	9.2G	81.6%	81.8%
Swin-S	4.6G	82.8%	82.9%
Swin-B	8.0G	83.1%	83.3%

1.4. Quantization results

Our analysis demonstrates that the ELSA framework, combined with quantization methods, maintains the inherent orthogonal characteristics between layer-wise sparsity and quantization compression. According to the results presented in Table 4, it is clear that the reduction in accuracy between compressed ELSA models before and after quantization is similar to that of the dense model before and after.

Table 4. Experimental results of the integration of ELSA with quantization technique (FQ-ViT [2]). Accuracy denotes the Top-1 accuracy measure on the ImageNet-1K validation set.

Model	FLOPs	Accuracy	
		FP32	INT8
DeiT-S	4.7G	79.82	79.47 (-0.35)
ELSA-DeiT-S-2:4	2.5G	79.14	78.86 (-0.28)
ELSA-DeiT-S-N:4	2.0G	78.34	77.77 (-0.57)
DeiT-B	17.6G	81.81	81.53 (-0.28)
ELSA-DeiT-B-2:4	9.2G	81.66	81.50 (-0.16)
ELSA-DeiT-B-N:4	6.0G	81.37	80.96 (-0.41)
Swin-S	8.7G	83.17	83.15 (-0.02)
ELSA-Swin-S-2:4	4.6G	82.81	82.68 (-0.13)
ELSA-Swin-S-N:4	3.5G	82.53	82.44 (-0.09)
Swin-B	15.4G	83.45	83.34 (-0.11)
ELSA-Swin-B-2:4	8.0G	83.10	83.14 (+0.04)
ELSA-Swin-B-N:4	5.3G	82.80	82.71 (-0.09)

References

[1] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. In *ECCV*, pages 544–560, 2020. 2

[2] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. FQ-ViT: Fully quantized vision transformer without retraining. *CoRR*, 2021. 3

[3] Asit K. Mishra, Jorge Albericio Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, and Paulius Micikevicius. Accelerating sparse deep neural networks. *CoRR*, abs/2104.08378, 2021. 3

[4] Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H Nguyen, Madeleine Gibescu, and Antonio Liotta. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature communications*, 9(1):2383, 2018. 1

[5] Wei Sun, Aojun Zhou, Sander Stuijk, Rob G. J. Wijnhoven, Andrew Nelson, Hongsheng Li, and Henk Corporaal. Dominosearch: Find layer-wise fine-grained n:m sparse schemes from dense neural networks. In *Advances in Neural Information Processing Systems*, 2021. 1