Variance-Based Pruning for Accelerating and Compressing Trained Networks

Supplementary Material

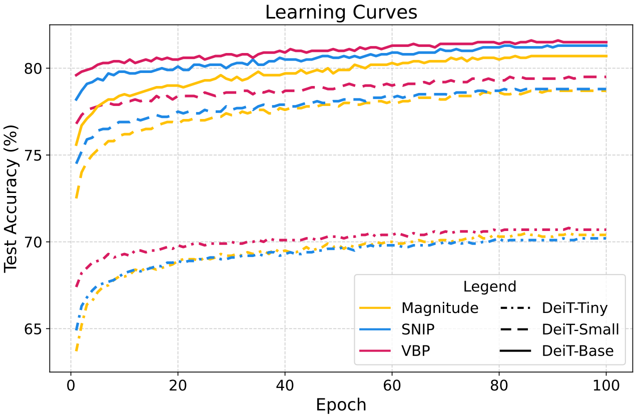

Figure 4. Learning curves over 100 epochs of fine-tuning after different structured pruning methods. VBP retains a performance lead over other methods throughout the entire training period.

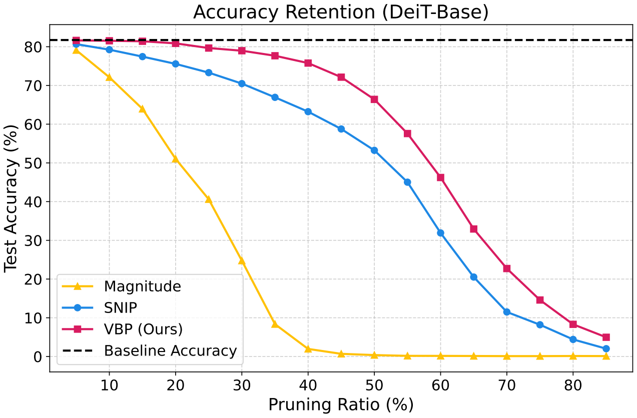

Figure 5. Accuracy retention across varying pruning ratios for different structured pruning methods applied to DeiT-Base. VBP consistently retains more accuracy across all pruning levels.

A. Additional Analysis

A.1. Longer Training and Learning Curves

To validate the sustained performance of VBP throughout training, we extend training to 100 epochs and compare learning curves against other pruning methods adapted for structured pruning. While the performance gap narrows over time, VBP consistently outperforms alternatives throughout all 100 epochs, as shown in Fig. 4.

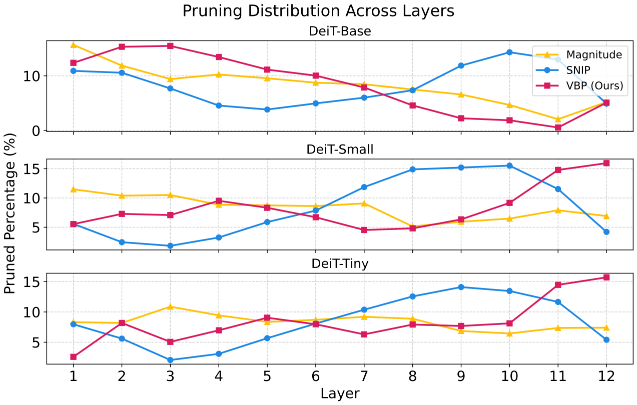

Figure 6. Layer-wise pruning distributions for different structured pruning methods. VBP prunes more in early layers, contrasting with SNIP which increases pruning in deeper layers.

A.2. Comparison at Different Pruning Ratios

We further examine the stability of VBP compared to other pruning methods adapted for structured pruning. We apply varying pruning rates to the DeiT-Base model and plot the accuracy retention. As shown in Fig. 5, VBP maintains a consistent advantage across all pruning levels, indicating its relative robustness to aggressive pruning.

A.3. Pruning Distribution Across Layers

We analyze how pruning decisions are distributed across layers. Interestingly, gradient-based methods such as SNIP tend to prune neurons in a pattern opposite to that of VBP, which focuses more on early layers. This difference is visualized in Fig. 6, suggesting that VBPs strategy may lead to better feature preservation and downstream performance.

B. Performance on ImageNet-1k for Different Pruning Rates

Model	**MACs (G)**	**Parameters (M)**	**Top-1 Acc. (%)**	
(Pruning Rate)			**Retention**	**Final**
DeiT-Base	17.58	86.57	-	81.73
5%	17.02 (-3.19%)	83.73 (-3.28%)	81.63 (99.91%)	81.79 (100.11%)
10%	16.46 (-6.37%)	80.90 (-6.55%)	81.54 (99.80%)	81.72 (100.02%)
15%	15.91 (-9.50%)	78.07 (-9.82%)	81.41 (99.65%)	81.73 (100.04%)
20%	15.35 (-12.68%)	75.24 (-13.09%)	80.87 (98.98%)	81.76 (100.07%)
25%	14.79 (-15.87%)	72.40 (-16.37%)	79.66 (97.50%)	81.76 (100.07%)
30%	14.23 (-19.06%)	69.57 (-19.64%)	78.97 (96.66%)	81.68 (99.98%)
35%	13.67 (-22.24%)	66.74 (-22.91%)	77.67 (95.07%)	81.71 (100.01%)
40%	13.12 (-25.37%)	63.90 (-26.19%)	75.78 (92.75%)	81.55 (99.82%)
45%	12.56 (-28.56%)	61.07 (-29.46%)	72.14 (88.30%)	81.32 (99.53%)
50%	12.00 (-31.74%)	58.24 (-32.72%)	66.40 (81.27%)	80.99 (99.13%)
DeiT-Small	4.61	22.05	-	79.70
5%	4.47 (-3.04%)	21.34 (-3.22%)	78.63 (98.66%)	79.67 (99.96%)
10%	4.33 (-6.07%)	20.63 (-6.44%)	78.32 (98.27%)	79.63 (99.91%)
15%	4.19 (-9.11%)	19.92 (-9.66%)	77.65 (97.43%)	79.65 (99.94%)
20%	4.05 (-12.15%)	19.22 (-12.83%)	76.90 (96.49%)	79.65 (99.94%)
25%	3.91 (-15.18%)	18.51 (-16.05%)	76.13 (95.52%)	79.63 (99.91%)
30%	3.77 (-18.22%)	17.80 (-19.27%)	74.98 (94.08%)	79.52 (99.77%)
35%	3.63 (-21.26%)	17.09 (-22.49%)	73.49 (92.21%)	79.30 (99.50%)
40%	3.49 (-24.30%)	16.38 (-25.71%)	71.44 (89.64%)	79.18 (99.35%)
45%	3.35 (-27.33%)	15.67 (-28.93%)	68.42 (85.85%)	78.90 (99.00%)
50%	3.21 (-30.37%)	14.96 (-32.15%)	64.44 (80.85%)	78.62 (98.64%)
DeiT-Tiny	1.26	5.72	-	72.02
5%	1.22 (-3.17%)	5.54 (-3.15%)	71.67 (99.54%)	72.05 (100.07%)
10%	1.19 (-5.56%)	5.36 (-6.29%)	70.95 (98.54%)	71.92 (99.89%)
15%	1.15 (-8.73%)	5.18 (-9.44%)	70.05 (97.29%)	71.76 (99.67%)
20%	1.12 (-11.11%)	5.01 (-12.41%)	68.87 (95.65%)	71.60 (99.44%)
25%	1.08 (-14.29%)	4.83 (-15.56%)	67.37 (93.57%)	71.44 (99.22%)
30%	1.05 (-16.67%)	4.65 (-18.71%)	64.76 (89.94%)	71.20 (98.89%)
35%	1.01 (-19.84%)	4.48 (-21.68%)	61.12 (84.89%)	70.86 (98.42%)
40%	0.978 (-22.38%)	4.30 (-24.83%)	55.64 (77.28%)	70.55 (97.99%)
45%	0.943 (-25.16%)	4.12 (-27.97%)	49.77 (69.13%)	70.08 (97.33%)
50%	0.908 (-27.94%)	3.94 (-31.12%)	39.58 (54.97%)	69.70 (96.81%)

Table 10. Comparison evaluating four metrics on ImageNet-1k [6]: **MACs**: computational operations, measured in billions of operations; and **Parameters**: the total model size in millions of parameters; **Accuracy Retention (Ret.)**: retained accuracy after pruning, before fine-tuning; and **Final Accuracy**: accuracy after fine-tuning. Our method achieves competitive accuracy with significant reductions in MACs and parameters and allows off-the-shelf deployment for pruning rates up to 20%.

C. Performance on CIFAR-100 for Different Pruning Rates

Model	**MACs (G)**	**Params (M)**	**Top-1 Acc. (%)**	
(Pruning Rate)			**Retention**	**Final**
DeiT-Base	17.58	86.57	-	88.23
30%	14.23 (-19.06%)	69.57 (-19.64%)	84.83 (96.15%)	88.07 (99.82%)
40%	13.12 (-25.37%)	63.90 (-26.19%)	82.21 (93.18%)	87.53 (99.21%)
50%	12.00 (-31.74%)	58.24 (-32.72%)	77.45 (87.78%)	87.00 (98.61%)
60%	10.88 (-38.11%)	52.57 (-39.27%)	54.60 (61.88%)	85.98 (97.45%)
70%	10.33 (-41.24%)	49.74 (-42.54%)	32.58 (36.93%)	84.84 (96.16%)
DeiT-Small	4.61	22.05	-	85.43
30%	3.77 (-18.22%)	17.80 (-19.27%)	71.88 (84.14%)	85.28 (99.82%)
40%	3.49 (-24.30%)	16.38 (-25.71%)	68.41 (80.08%)	85.48 (100.06%)
50%	3.21 (-30.37%)	14.96 (-32.15%)	60.99 (71.39%)	84.94 (99.43%)
60%	2.93 (-36.44%)	13.55 (-38.55%)	51.17 (59.90%)	83.65 (97.92%)
70%	2.65 (-42.52%)	12.13 (-44.99%)	33.19 (38.85%)	82.08 (96.08%)
DeiT-Tiny	1.26	5.72	-	80.50
30%	1.05 (-16.67%)	4.65 (-18.71%)	70.94 (88.12%)	80.15 (99.57%)
40%	0.98 (-22.22%)	4.30 (-24.83%)	62.10 (77.14%)	79.97 (99.34%)
50%	0.91 (-27.78%)	3.94 (-31.12%)	46.01 (57.16%)	78.52 (97.54%)
60%	0.84 (-33.33%)	3.59 (-37.24%)	24.68 (30.66%)	77.20 (95.90%)
70%	0.77 (-38.89%)	3.23 (-43.53%)	9.33 (11.59%)	74.11 (92.06%)

Table 11. Comparison evaluating four metrics on CIFAR-100 [11]: **MACs**: computational operations, measured in billions of operations; **Parameters**: the total model size in millions of parameters; **Accuracy Retention (Ret.)**: retained accuracy after pruning, before fine-tuning; and **Final Accuracy**: accuracy after fine-tuning.

D. Runtime Performance for Reported Models

Model	H200 (GPU)			T4 (GPU)			E5 (CPU)		
(Pruning Rate)	Time (ms)		Speed	Time (ms)		Speed	Time (ms)		Speed
	Full	VBP	Up	Full	VBP	Up	Full	VBP	Up
DeiT-T (45%)	3.64ms	3.10ms	1.17×	33.17ms	28.36ms	1.17×	0.53s	0.43s	1.25×
DeiT-S (50%)	9.81ms	7.34ms	1.34×	107.42ms	81.05ms	1.33×	1.72s	1.20s	1.43×
DeiT-B (55%)	30.73ms	21.36ms	1.44×	378.63ms	273.94ms	1.38×	5.71s	3.81s	1.50×
DeiT-B (20%)	30.73ms	27.72ms	1.11×	378.63ms	342.34ms	1.11×	5.71s	5.12s	1.11×
Swin-T (45%)	14.00ms	11.66ms	1.20×	131.78ms	111.00ms	1.19×	2.70s	2.28s	1.18×
Swin-S (50%)	24.82ms	19.60ms	1.27×	241.27ms	186.39ms	1.29×	4.91s	3.76s	1.30×
Swin-B (55%)	36.71ms	28.14ms	1.30×	376.82ms	288.13ms	1.31×	7.52s	5.75s	1.31×
Swin-B (20%)	36.71ms	33.37ms	1.10×	376.82ms	341.97ms	1.10×	7.52s	6.85s	1.10×
ConvNeXt-T (45%)	12.07ms	9.43ms	1.28×	133.20ms	106.44ms	1.25×	1.93s	1.48s	1.30×
ConvNeXt-S (50%)	21.62ms	15.20ms	1.42×	248.75ms	172.96ms	1.44×	3.37s	2.16s	1.55×
ConvNeXt-B (55%)	32.39ms	21.75ms	1.49×	384.87ms	249.95ms	1.54×	5.64s	3.30s	1.71×
ConvNeXt-B (20%)	32.39ms	28.49ms	1.14×	384.87ms	340.90ms	1.13×	5.64s	4.91s	1.15×

Table 12. Comparison of runtime performance across different hardware: **H200 (GPU)**: high-end NVIDIA Tensor Core GPUs; **T4 (GPU)**: cost-efficient NVIDIA Tesla GPUs for inference; and **E5 (CPU)**: Intel Xeon E5-2680v4 processors. We report **Full** and **VBP** runtimes in milliseconds, along with the resulting **Speed-up** factors. Our method consistently improves inference latency across diverse hardware environments, reaching up to 1.71× speed-ups.

E. Activation Distributions Across Layers

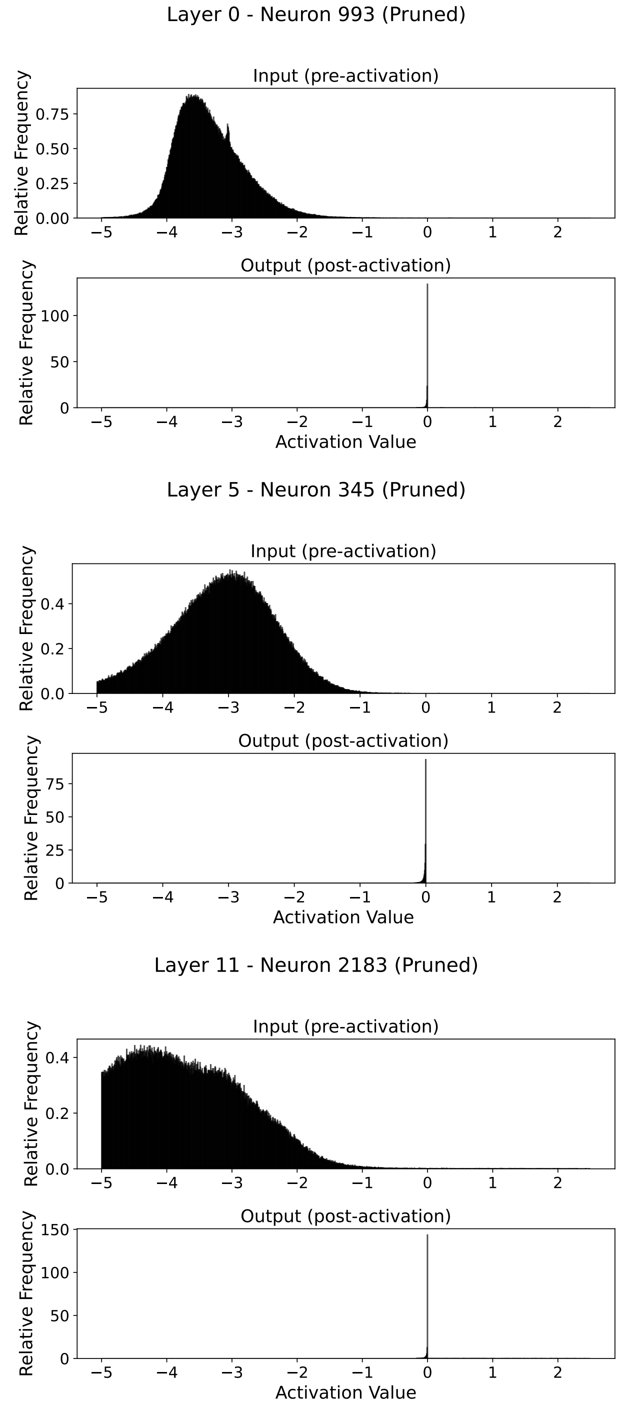

Figure 7. Visualization of activation distributions before and after the non-linearity in pruned neurons throughout different layers.

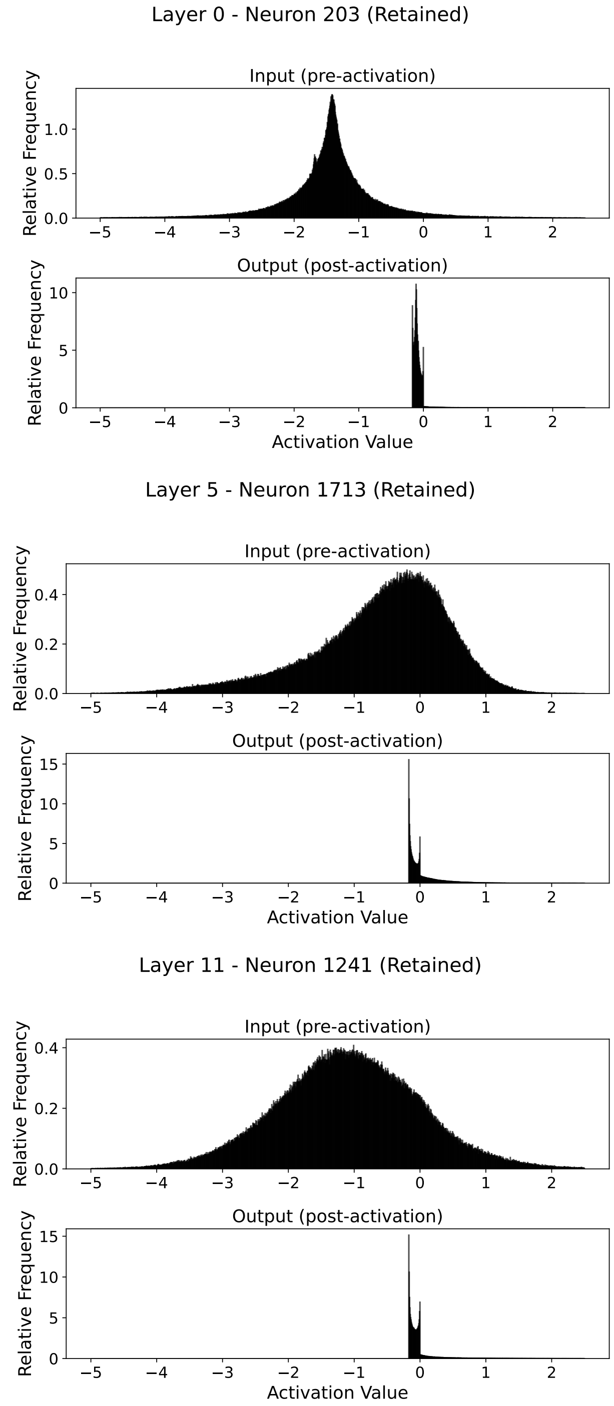

Figure 8. Visualization of activation distributions before and after the non-linearity in retained neurons throughout different layers.