SAFER: Sharpness Aware layer-selective Finetuning for Enhanced Robustness in vision transformers

Supplementary Material

A. Top-sharpness layers by model and dataset

Table 8 lists the indices of layers exhibiting the highest sharpness in a PGD-AT pretrained ViT model. Notably, while sharp layers vary across models, the dataset used for training and evaluation has minimal effect on the sharpness ranking of layers. This aligns with our observation in Table 7, where sharpness rankings remain consistent across different random evaluation batches. This consistency suggests that certain layers have inherent structural properties, predisposing them to overfitting during adversarial training. A theoretical exploration of this observation is reserved for future work. Theoretical analysis on this observation is left for future work.

Table 8. Indices of the Top-5 sharpest layers by model and dataset. Layers selected in SAFER finetuning are bolded.

MODEL	CIFAR10	CIFAR100
DEIT-TI	**11, 10**, 13, 8, 16	**11, 10**, 13, 8, 16
VIT-S	**5, 9**, 14, 20, 25	**5, 9**, 15, 20, 25
SWIN-B	**4, 15, 32, 46**, 50	**4, 16, 32, 46**, 50

B. Ablation: Dynamic layer selection

During the SAFER finetuning process, we recompute the sharpness measurement every 10 epochs to update our selection of layers most susceptible to overfitting. Table 9 shows that dynamically selecting layers for finetuning is crucial to SAFER's performance. In contrast, fixing the same set of layers that are selected on the initial pretrained model throughout finetuning results in significantly lower accuracies compared to the baselines. Finetuning in this study was conducted for 20 epochs. Although not shown, the performance gap becomes more pronounced with extended finetuning, as the initially selected layers become less prone to overfitting, providing minimal improvements with further tuning.

C. Ablation: Number of layers chosen

Figure 3 illustrates the performance variation between adversarial accuracy and the number of layers selected for SAFER finetuning in both DeiT-Ti and ViT-S. Initially, increasing the number of layers improves model flexibility, resulting in enhanced SAFER performance. However, beyond a certain point, finetuning additional layers reduces SAFER's effectiveness, likely due to the model focusing on less-relevant layers, which complicates convergence under the SAM objective. This trend is consistent across datasets, with results shown for Imagenette and CIFAR-10.

Table 9. Ablation study on dynamic and fixed sharp layer selection for SAFER on CIFAR-10: The columns present clean, PGD-20 and Auto Attack (AA) evaluation accuracies for models trained with SAFER, using dynamic or fixed layers for fine-tuning.

NETWORK	DYNAMIC LAYERS			FIXED LAYERS		
	CLEAN	PGD-20	AA	CLEAN	PGD-20	AA
DEIT-TI	82.36	68.50	50.12	80.10	64.49	47.53
VIT-S	83.40	68.89	50.12	79.92	63.12	46.21
SWIN-B	86.52	53.65	52.00	84.39	50.45	50.19

Both models identify selecting the Top-2 layers (approximately 5% of the 36 total layer options) as leading to the best results. As ViT-S layers are significantly larger, selecting additional layers for ViT-S results in a steeper decline in performance, primarily due to optimization difficulties arising from the increased parameter count. This observation highlights the importance of selecting the optimal number of layers during SAFER finetuning.

D. Additional attacks for robustness evaluation

D.1. Convergence of the PGD attack

In the main paper, we report PGD attack robustness using attacks with 20 gradient ascent steps. As suggested by **?**], PGD attacks with insufficient update steps may be ineffective due to gradient masking, resulting in inaccurate robustness measurements. To address this, Tab. 10 presents the robustness results under PGD attacks with increased steps for selected models reported in Table 1. As shown in the table, increasing the attack steps does not result in further decreases in model robustness. This demonstrates that the adversarial images generated in the main paper originate from well-converged attacks, and adding more steps does not improve convergence.

D.2. Additional attack types

To further demonstrate SAFER's effectiveness in improving model robustness, we compare models trained with SAFER to those trained with PGD-AT (SAM) on the CIFAR-10 dataset, as reported in Table 1. The evaluation includes stronger white-box attacks that are not limited to ℓ_∞-bounded constraints. Table 11 shows the robustness re-

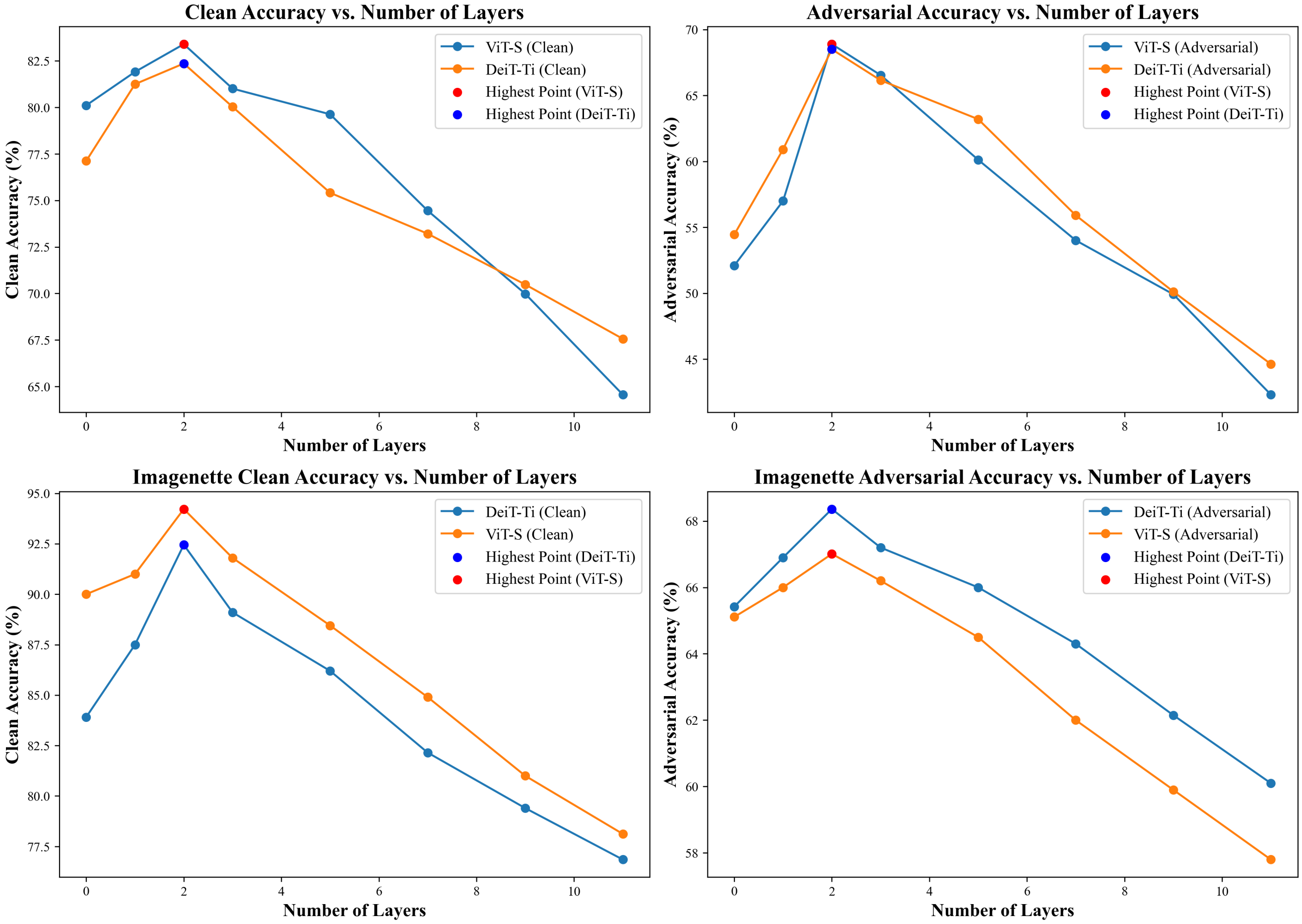

Figure 3. Performance comparison of DeiT-Ti and ViT-S as a function of number of sharp layers selected for SAFER finetuning. The top row shows CIFAR-10 clean and adversarial accuracy, while the bottom row shows Imagenette results. The number “0” on the X axis corresponds to PGD-AT (SAM) without SAFER, where fine-tuning is performed on the entire model. The highest performance points are highlighted in blue for DeiT-Ti and red for ViT-S.

Table 10. SAFER adversarial accuracy on CIFAR-10 under PGD attacks with 20, 50, and 100 steps

MODEL	PGD-20 (%)	PGD-50 (%)	PGD-100 (%)
DEIT-TI	68.50	68.12	68.04
VIT-S	68.89	68.51	68.73
CONVIT-B	56.21	56.34	56.22
SWIN-B	53.65	53.22	53.05

sults under the FAB attack [?], StAdv attack [?], PIXEL attack [?], ℓ_∞-bounded PGD attacks with higher strengths, and ℓ_2-bounded PGD attack. Across all evaluated attacks, SAFER-trained models consistently demonstrate robustness improvements over baseline models.

E. Learning curve under extended training

It has been observed that adversarial overfitting can be mitigated by early stopping [?]. SAFER is designed to eliminate the need for early stopping and fully leverage the model’s learning potential throughout the finetuning process. To this end, we extend the learning curve experiments in Figure 2 to 150 adversarial training epochs. A cosine learning rate scheduler is used so that the learning rate decays to 0 by epoch 150. As shown in Figure 4, even with the SAM optimizer, models trained with PGD-AT show a consistent decline in performance with additional epochs of adversarial training, highlighting the effects of overfitting. In contrast, models trained with SAFER (whether pretrained model or from scratch), show consistent performance and robustness improvement throughout all epochs. This further proves the effectiveness of SAFER in countering overfitting.

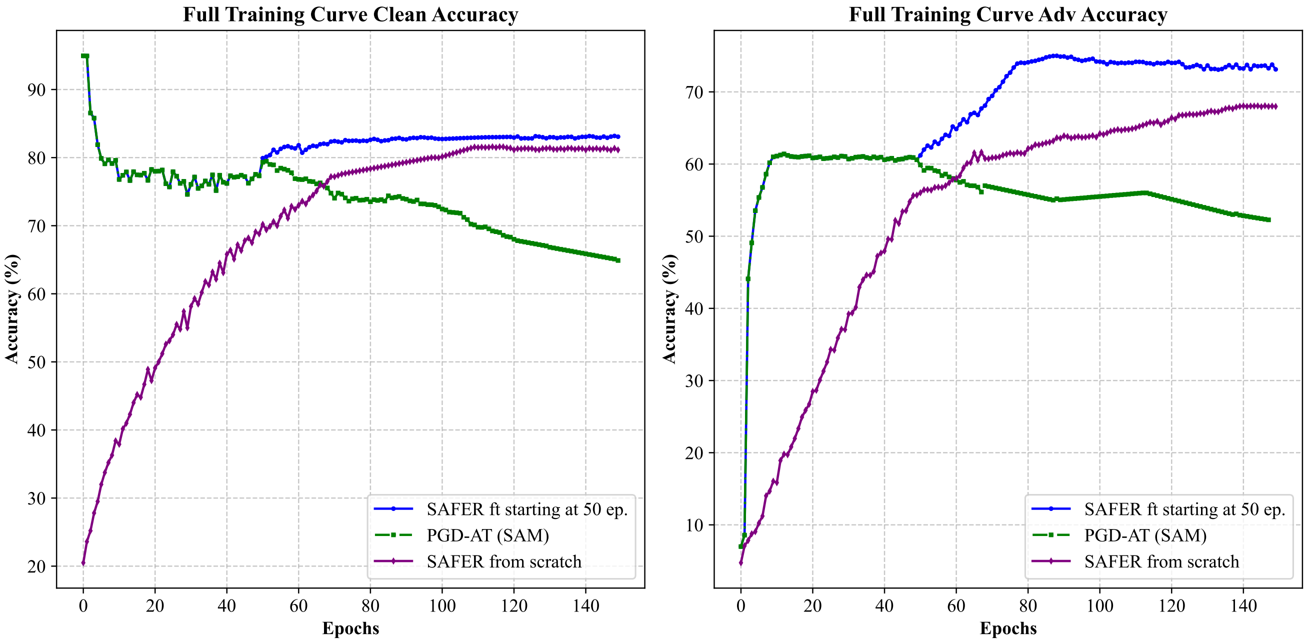

Figure 4. SAFER vs. PGD-AT (SAM) performance on CIFAR-10 with DeiT-Ti: clean (left) vs. adversarial (right) accuracies.

Table 11. Adversarial accuracies across various attacks on CIFAR-10, comparing models trained on DeiT-Ti and ViT-S without and with SAFER training/finetuning, respectively. Positive blue values indicate performance improvements achieved with SAFER-trained models over PGD-AT (SAM) baseline.

Attack	Method	DeiT-Ti (%)	ViT-S (%)
FAB	PGD-AT (SAM)	24.79	26.52
	SAFER	+3.36	+2.61
StAdv	PGD-AT (SAM)	19.60	20.21
	SAFER	+3.85	+4.54
Pixel	PGD-AT (SAM)	7.30	8.40
	SAFER	+1.40	+1.50
PGD-20 L_∞	PGD-AT (SAM)	54.45	52.10
($\epsilon = 0.03$)	SAFER	+14.05	+16.79
PGD-20 L_∞	PGD-AT (SAM)	47.20	46.28
($\epsilon = 0.05$)	SAFER	+6.59	+8.63
PGD-20 L_∞	PGD-AT (SAM)	40.25	41.79
($\epsilon = 0.07$)	SAFER	+9.94	+10.03
PGD-20 L_2	PGD-AT (SAM)	56.79	56.05
($\epsilon = 0.03$)	SAFER	+12.33	+14.13