On the Compatibility of Adversarial Training and Sparse Training in Vision Transformer

Zhenting Hu, Jiawang Xu, Haihan Zhang

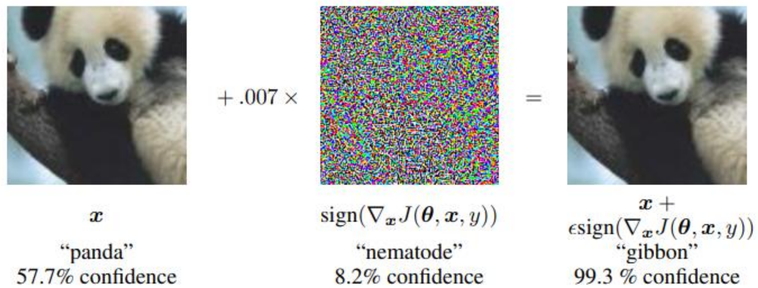
Background:Data Security on Edge Devices

- Edge Devices Hold Sensitive Raw Data
- Limited Defense Capabilities

Task: Build a defense model that runs directly on the edge devices with good privacy and robustness.

Stronger Robustness: Adversarial Training

Adversarial Examples

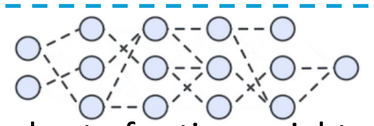


Attack Method:FGSM(Fast Gradient Sign Method)

Why improve robustness?

Training models on both clean and adversarial samples

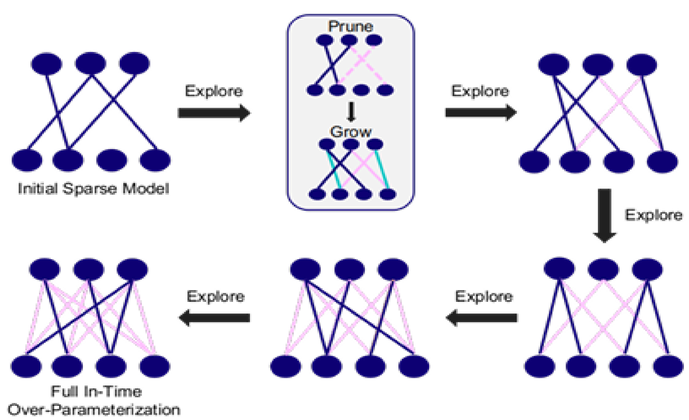
Smaller Model: Sparse Training



Why smaller size?

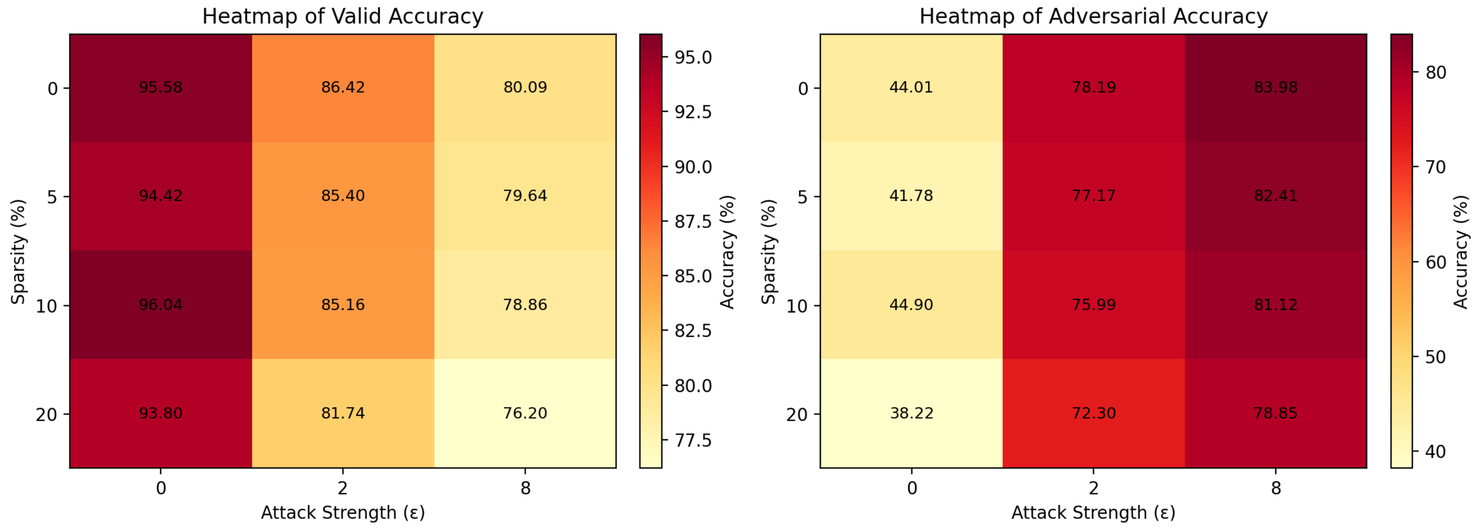
Training neural networks with only a subset of active weights

In-Time Over-Parameterization in Sparse Training

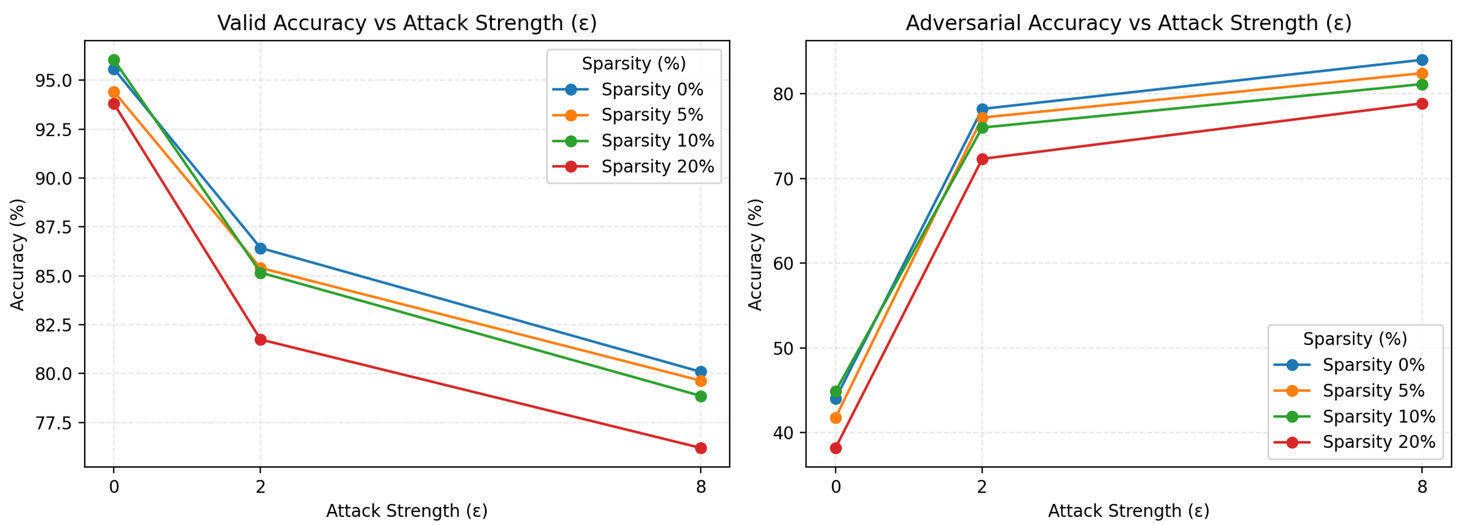


Dataset:CIFAR-10, Model: Vision Transformer

Accuracy Heatmap: Sparsity × ε



Robustness Comparison under Different Sparsity Levels



Conclusion:

- Clean accuracy decreases as ε increases, with sparsity having little influence.
- Adversarial accuracy improves at higher ε values.
- High sparsity (20%) weakens robustness, while moderate sparsity (5–10%) provides the best balance