Assignment 1 - White-Box Attack (FGSM)

CAP6938 - Trustworthy Machine Learning - Alexander Green

September 2025

1 Methodology

1.1 Datasets and Models

I evaluate my implementation on two benchmark datasets: MNIST (28×28 grayscale digits, 10 classes) and CIFAR-10 (32×32 color images, 10 classes).

Model	Architecture	MNIST Configuration	CIFAR-10 Configuration
ResNet-18	Convolutional with residual connections	Input channels: 1 4 residual blocks: 64→128→256→512 BatchNorm + ReLU	Input channels: 3 4 residual blocks: 64→128→256→512 BatchNorm + ReLU
ViT	Vision Transformer with self-attention	Patch size: 4×4 (49 patches) Embed dim: 256, 8 layers, 4 heads CLS token + position embeddings	Patch size: 4×4 (64 patches) Embed dim: 256, 8 layers, 4 heads CLS token + position embeddings

1.2 Training Configuration

Parameter	ResNet-18	ViT
Optimizer	Adam (lr=0.001, weight_decay=$1e^{-4}$)	AdamW (lr=0.00025, weight_decay=0.05)
Scheduler	StepLR (step=5, $\gamma = 0.1$)	CosineAnnealingLR
Epochs	MNIST: 10, CIFAR-10: 30	MNIST: 10, CIFAR-10: 30
Special Features	Standard training	Gradient clipping (max_norm=1.0) Early stopping (patience=5) Extra dropout (0.3)
Data Augmentation	CIFAR-10: RandomCrop + HorizontalFlip	CIFAR-10: RandomCrop + HorizontalFlip

1.3 FGSM Implementation

Untargeted Attack:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y_{\text{true}}))$$

Targeted Attack:

$$x_{\text{adv}} = x - \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y_{\text{target}}))$$

Where x is the original input, ϵ the perturbation budget, L the cross-entropy loss, y_{true} the true label, and y_{target} the target label.

2 Results

This section reports baseline performance, adversarial robustness results, and all visualizations generated by the implementation. Figures 1–6 were produced automatically by `main.py`.

2.1 Baseline Model Performance

Model	Dataset	Clean Accuracy	Training Epochs	Final Training Acc
ResNet-18	MNIST	99.89%	10	99.90%
ResNet-18	CIFAR-10	91.33%	30	91.33%
ViT	MNIST	98.40%	10	98.40%
ViT	CIFAR-10	82.54%	30	82.54%

2.2 Adversarial Robustness Results

All values reported in the following tables are percentages (%).

2.2.1 MNIST Results

Model	Clean Acc	Robust 1/255	Robust 2/255	Robust 4/255	Robust 8/255	ASR Untargeted 8/255	ASR Random Target 8/255	ASR Least-Likely 8/255
ResNet-18	99.7	9.7	9.7	9.7	9.7	90.3	10.2	0.0
ViT	98.6	33.0	33.1	33.1	32.9	67.1	6.2	0.2

2.2.2 CIFAR-10 Results

Model	Clean Acc	Robust 1/255	Robust 2/255	Robust 4/255	Robust 8/255	ASR Untargeted 8/255	ASR Random Target 8/255	ASR Least-Likely 8/255
ResNet-18	88.4	44.5	42.2	38.4	31.7	68.3	10.6	1.2
ViT	79.8	35.0	34.2	31.7	28.6	71.4	9.5	1.5

2.3 Attack Visualizations

Note that all figures are present in the ”results” folder of my submission. If text is too small to read, please feel free to open the images there.

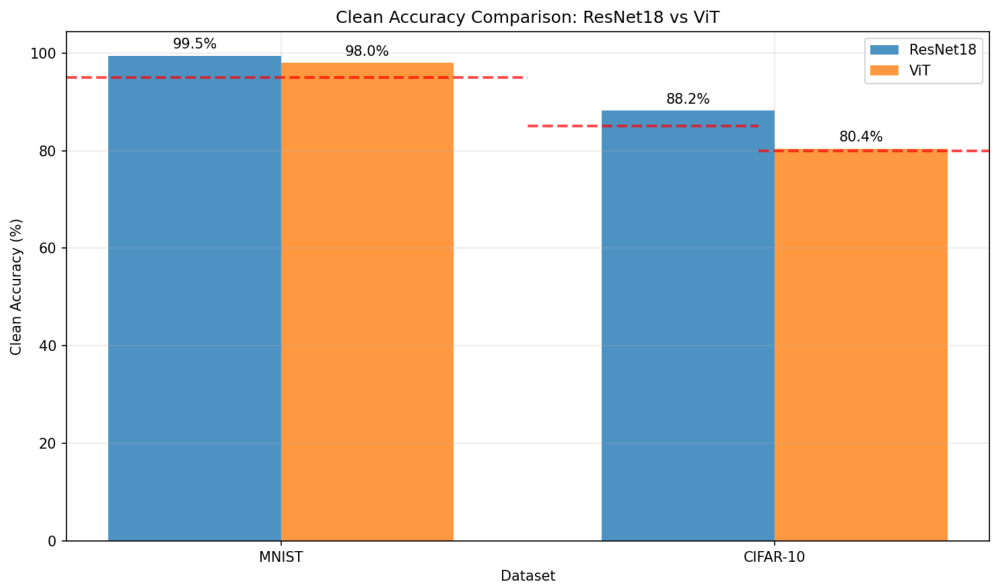

Figure 1: Clean Accuracy Comparison: ResNet-18 vs ViT.

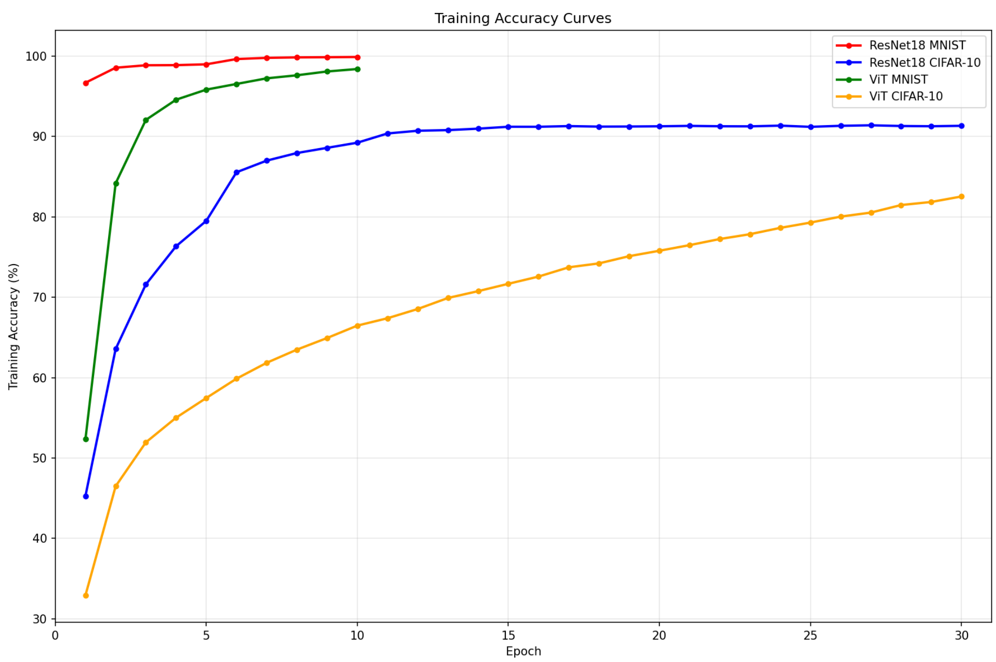

Figure 2: Training Accuracy Curves for both models.

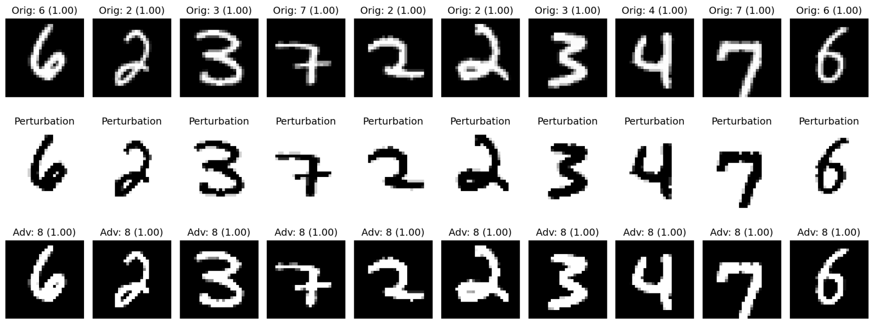

Figure 3: FGSM Attack Visualizations - ResNet-18 MNIST ($\epsilon = 8/255$).

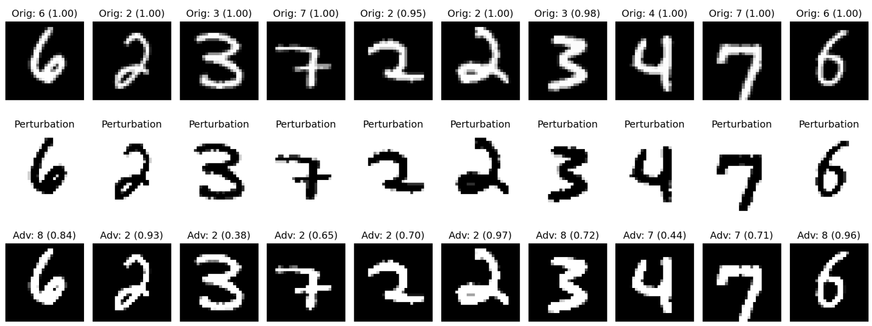

Figure 4: FGSM Attack Visualizations - ViT MNIST ($\epsilon = 8/255$).

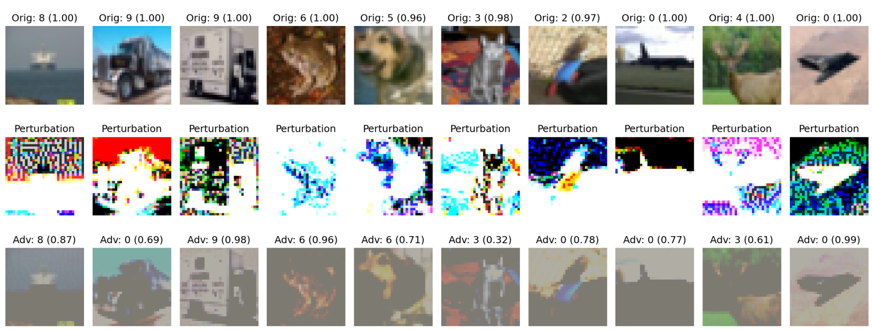

Figure 5: FGSM Attack Visualizations - ResNet-18 CIFAR-10 ($\epsilon = 8/255$).

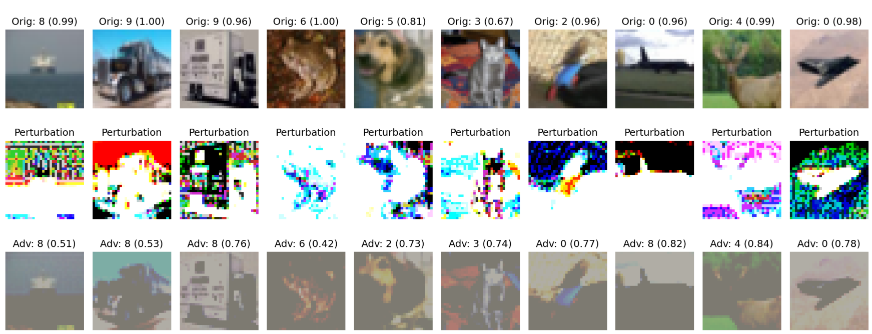

Figure 6: FGSM Attack Visualizations - ViT CIFAR-10 ($\epsilon = 8/255$).

3 Analysis

3.1 Architecture Comparison

ResNet-18 maintains higher clean accuracy than ViT across both datasets. However, the results reveal that for MNIST, ResNet-18 is almost completely compromised under FGSM attacks at even the smallest perturbation budget ($\epsilon = 1/255$), with robust accuracy dropping to 9.7%, whereas ViT retains about 33% robust accuracy. On CIFAR-10, ResNet-18 retains slightly higher clean accuracy (88.4% vs 79.8%) and modestly better robustness at mid-range ϵ values, though both architectures ultimately exhibit substantial vulnerability. These findings indicate that while convolutional inductive biases improve baseline accuracy, they do not confer meaningful resistance to FGSM attacks, especially on simpler datasets like MNIST.

3.2 Targeted vs Untargeted Attack Behavior

The hierarchy of attack success remains consistent: untargeted attacks achieve the highest ASR, followed by random and least-likely targeted attacks. For MNIST, untargeted ASR against ResNet-18 reaches 90.3%, while ViT experiences 67.1% ASR. Least-likely targeted attacks remain largely ineffective on MNIST for both models. On CIFAR-10, untargeted attacks are highly effective for both architectures (ResNet-18 ASR: 68.3%, ViT ASR: 71.4%), indicating that increased visual complexity does not substantially alter the relative difficulty of targeted attacks compared to untargeted attacks.

3.3 Dataset-Specific Patterns

MNIST exhibits stark vulnerability despite its simplicity, with ResNet-18 nearly failing under FGSM and ViT retaining moderate robustness. CIFAR-10 models show lower clean accuracy and slightly better mid-range robustness for ResNet-18, reflecting that dataset complexity interacts with both baseline performance and attack susceptibility. Overall, CIFAR-10 models are more fragile in practical terms, with robust accuracy dropping below 32% for both architectures at $\epsilon = 8/255$.

3.4 Notable Failure Cases and Sensitivities

Robustness degrades precipitously with increasing ϵ. For MNIST, even $\epsilon = 1/255$ drastically reduces ResNet-18 performance, highlighting extreme sensitivity to small perturbations. ViT shows a more gradual degradation, indicating its self-attention mechanism may offer limited smoothing of gradients. On CIFAR-10, both models degrade sharply between $\epsilon = 4/255$ and $\epsilon = 8/255$, with ASR exceeding 68% for untargeted attacks. The results underscore that FGSM is highly effective at revealing intrinsic model weaknesses, with vulnerability largely independent of architecture but influenced by baseline accuracy and dataset characteristics.