Training Robust Neural Networks

attaqueoudefense

Victor Lemaître, Azammat Charaf Zadah, Louis Carron

IASD - Université PSL - Paris Dauphine

Academic Year 2025-2026

Outline

1. Setup

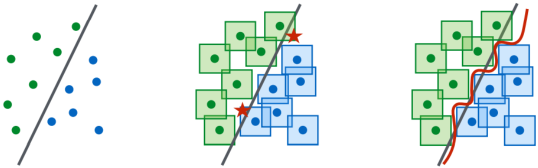
- Implemented attacks FGSM & PGD
- Standard adversarial training:
  - Perturb a fraction of each batch
  - Trade-off: clean accuracy drops too much

Key idea: Curriculum Adversarial Training

$$\begin{cases} t = \dfrac{\text{epoch}}{\text{total epochs}} \\ \epsilon(t) = 0.01(1-t) + 0.1\,t \\ \text{PGD iterations}(t) = \mathrm{int}\big(10(1-t) + 40\,t\big) \end{cases}$$

2. Improving Model Capacities

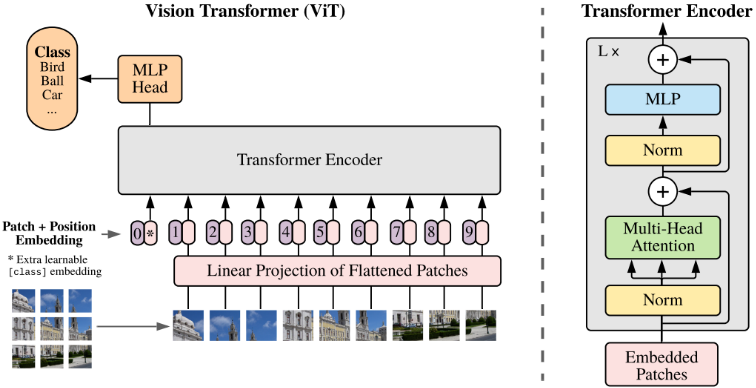
Key idea: stronger models improve the clean/attacked trade-off.



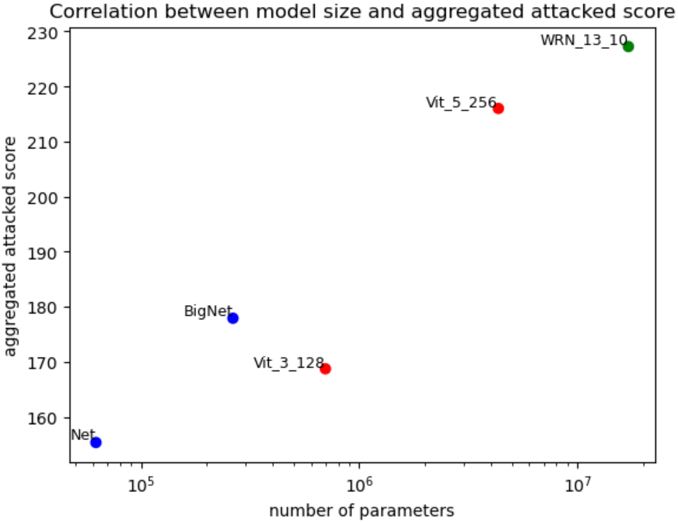
Practical improvements:

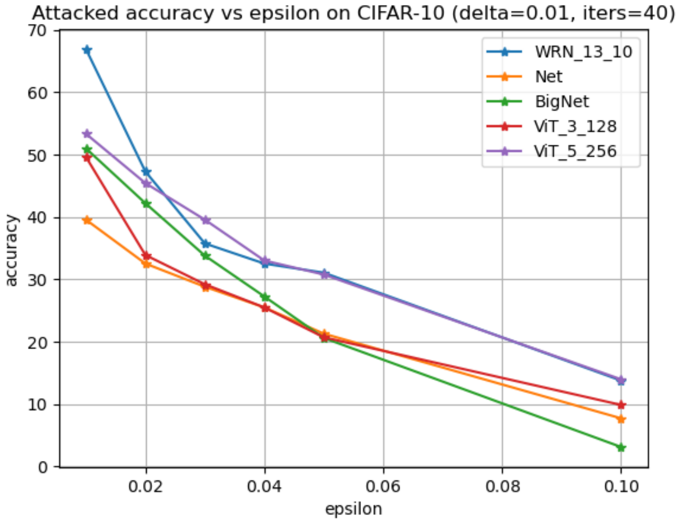
- Switch from SGD+momentum to AdamW
- Use GELU activations (+5–7% natural accuracy on larger models)

2. Improving Model Capacities

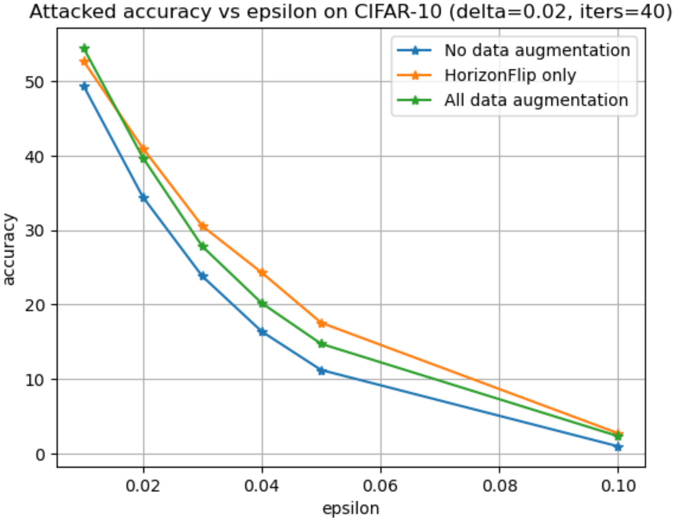


2.1 Model Size



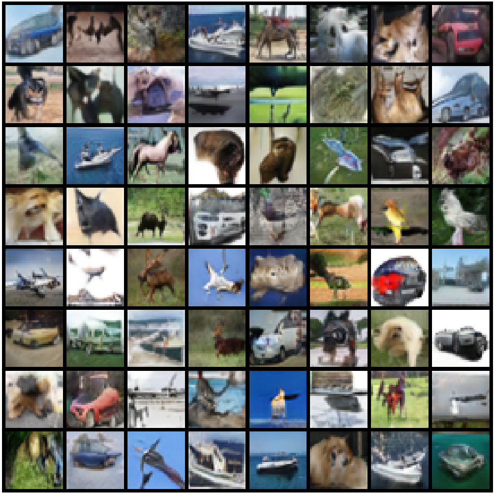

Attacked accuracy vs epsilon on CIFAR-10 (delta=0.01, iters=40)
accuracy
epsilon
0
10
20
30
40
50
60
70
0.02
0.04
0.06
0.08
0.10
WRN_13_10
Net
BigNet
ViT_3_128
ViT_5_256

2.2 Data Augmentation



Impact of data augmentation on BiggerNet (horizontal flip + random cropping + color jittering)

2.2 Synthetic data



The best we could get with a large conditional GAN, spectral normalization on the discriminator, Hinge loss, and exponential moving average of the generator's weights.

3. Randomness

Key idea: explore stochastic defenses to improve robustness.

Methods implemented:

- **Dropout**
- **PNI — Parametric Noise Injection**
  - Gaussian noise added to weight tensor
  - Learnable scaling factor α_l

$$\tilde{v}_{l,i} = v_{l,i} + \alpha_l \, \eta_{l,i}, \qquad \eta_{l,i} \sim \mathcal{N}(0, \sigma^2)$$

- **RSE — Random Self-Ensemble**
  - Gaussian noise layer before each convolution
  - Two fixed noise levels: σ_{init} and σ_{inner}
  - Noise parameters are *not* learned

3. Randomness — Experimental Results

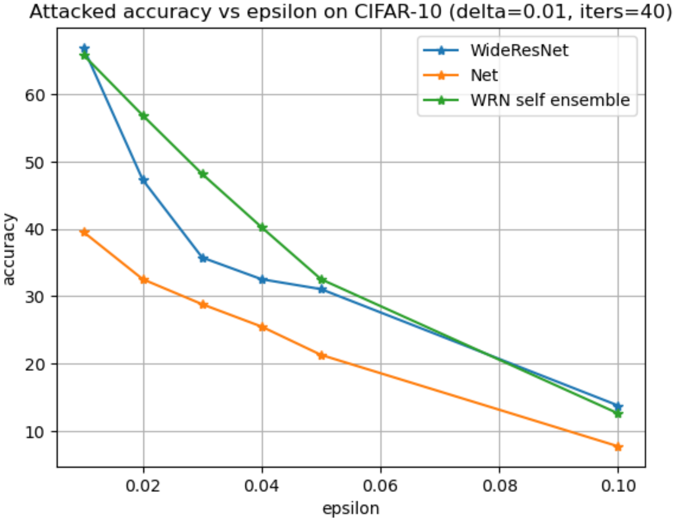
PGD-based adversarial training:

Model	**BigNet**	**BigNet-PNI**	**BigNet-RSE**
Natural acc.	53.32%	51.66%	49.41%
Attacked acc.	31.15%	31.07%	**32.50%**

Results of adversarial training for BigNet and randomized defense variants. (10 epochs, $\epsilon = 0.03$, $\delta = 0.008$)

- Small scale experiments first
- PNI: no significant improvement
- RSE: best attacked accuracy, slight clean accuracy drop

3. Randomness — Experimental Results



Base Net, WRN, WRN-RSE against PGD attacks (no data augmentation)

4. Towards Black Box attack

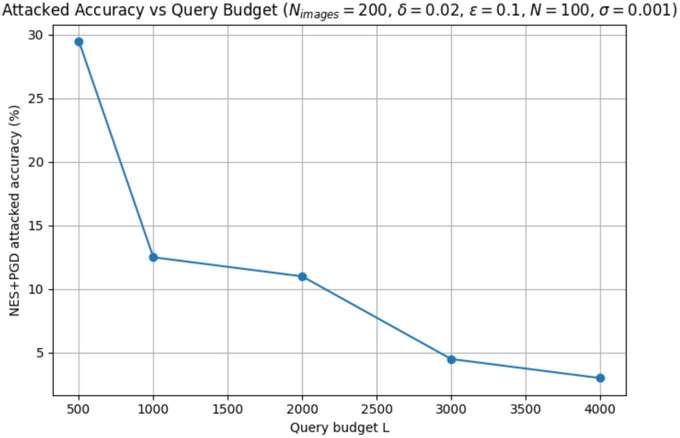
Setting: True gradients are not accessible.

Method: NES Gradient Estimator

$$\hat{\nabla} P(y|x) = \frac{1}{2n\sigma} \sum_{i=1}^{n} \left[P(y|x + \sigma u_i) - P(y|x - \sigma u_i) \right] u_i .$$

- P(y|x) classifier probability for class y given x
- $u_i \sim \mathcal{N}(0, I)$ are isotropic Gaussian directions
- σ the search variance.
- Given a total query budget L and a gradient budget N, perform $\frac{L}{N}$ steps of PGD.

4. Towards Black Box attack



5. Other Failed Improvement

- **Spectral normalization**
- **Low-rank PCA projection**
- **Gaussian noise on inputs**