# Characterising Robustness via Natural Input Gradients

Adrian Rodriguez-Muñoz, Tongzhou Wang, Antonio Torralba

EUROPEAN CONFERENCE ON COMPUTER VISION

MILAN

Adversarial Training is SOTA but **expensive** (minimax)

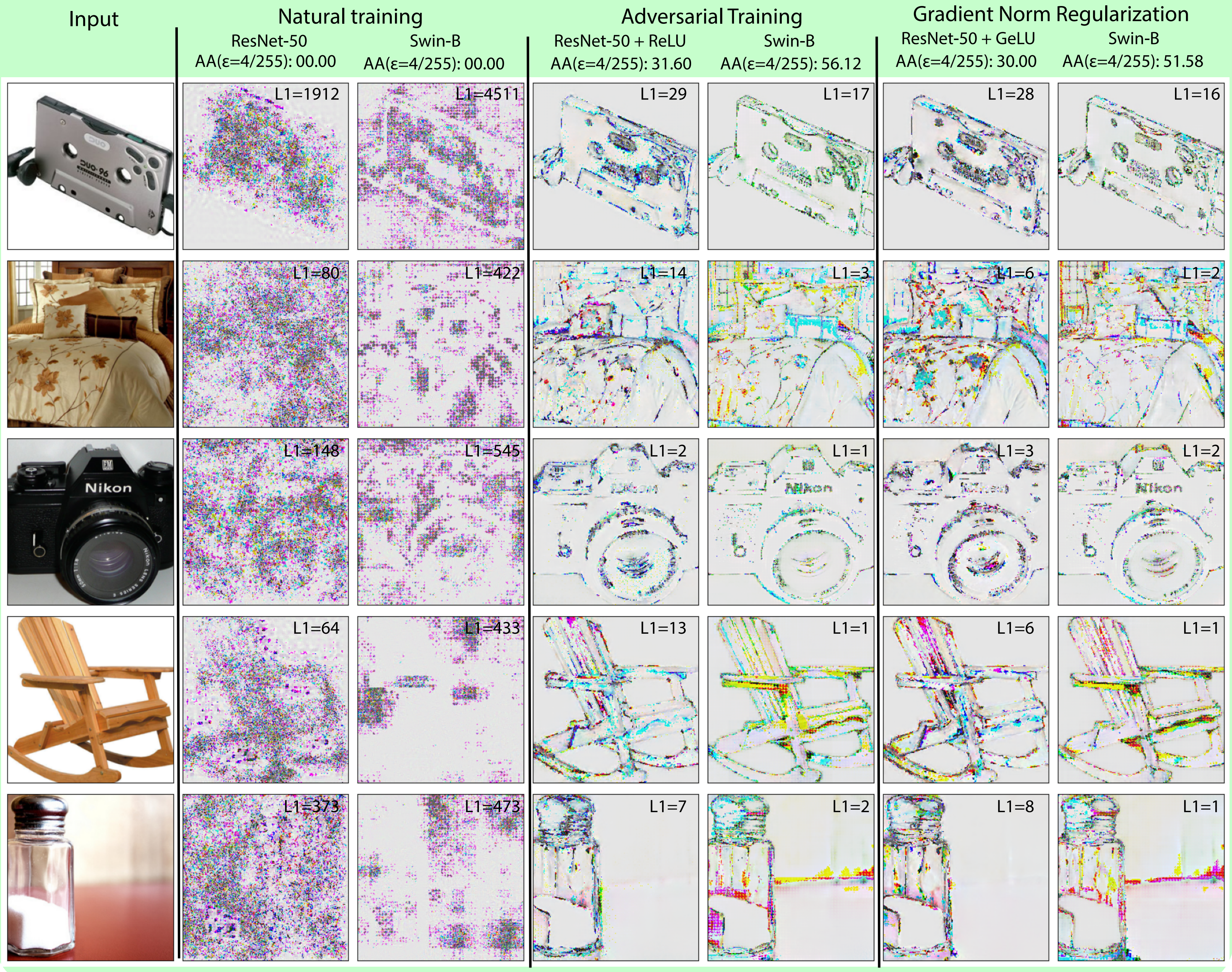
We get **>90% of the performance** at **60% of the cost** via **gradient regularisation** on ImageNet

Key ingredients for large scale: **smooth activation** & **adaptive optimisers**

## Definition of loss-input gradients

$$\nabla_x \mathcal{L} := \underbrace{\nabla_x \mathcal{L}_{\text{CE}}(f_\theta(x), y)}_{\text{loss-input gradient of model } f_\theta \text{ on example } x \text{ with groundtruth class } y},$$

## Visualisation of loss-input gradients




## Definition of GradNorm loss

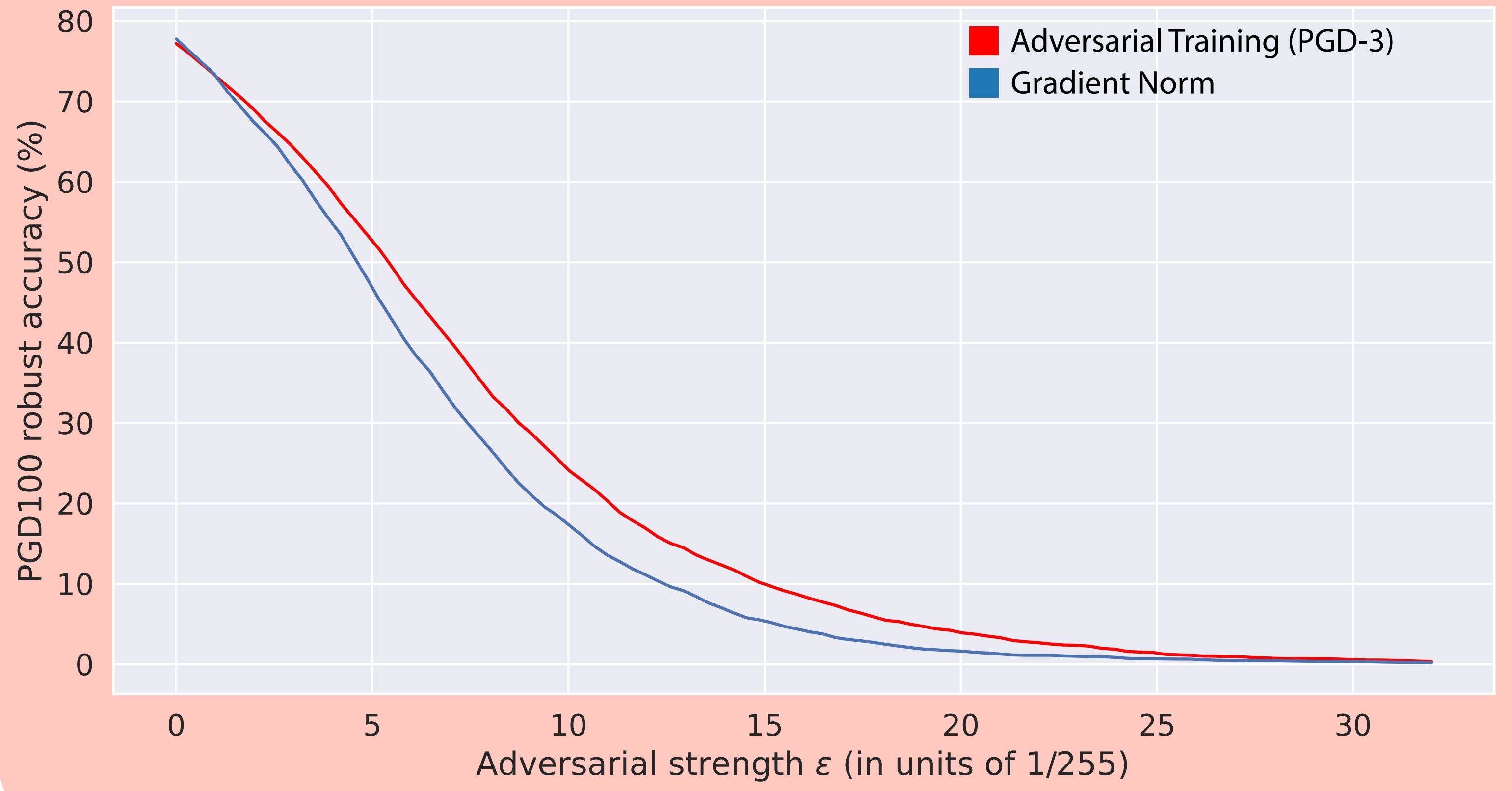
$$\mathcal{L}_{\text{GradNorm}}(\mathbf{x}, y) = \lambda_{\text{CE}}\mathcal{L}_{\text{CE}}(f_\theta(\mathbf{x}), y) + \lambda_{\text{GN}}\frac{\epsilon}{\sigma}||\nabla_{\mathbf{x}}\mathcal{L}_{\text{CE}}(f_\theta(\mathbf{x}), y))||_1$$

Where $\lambda_{CE} = 0.8$, $\lambda_{GN} = 1.2$, $\epsilon = 4/255$, $\sigma = 0.224$

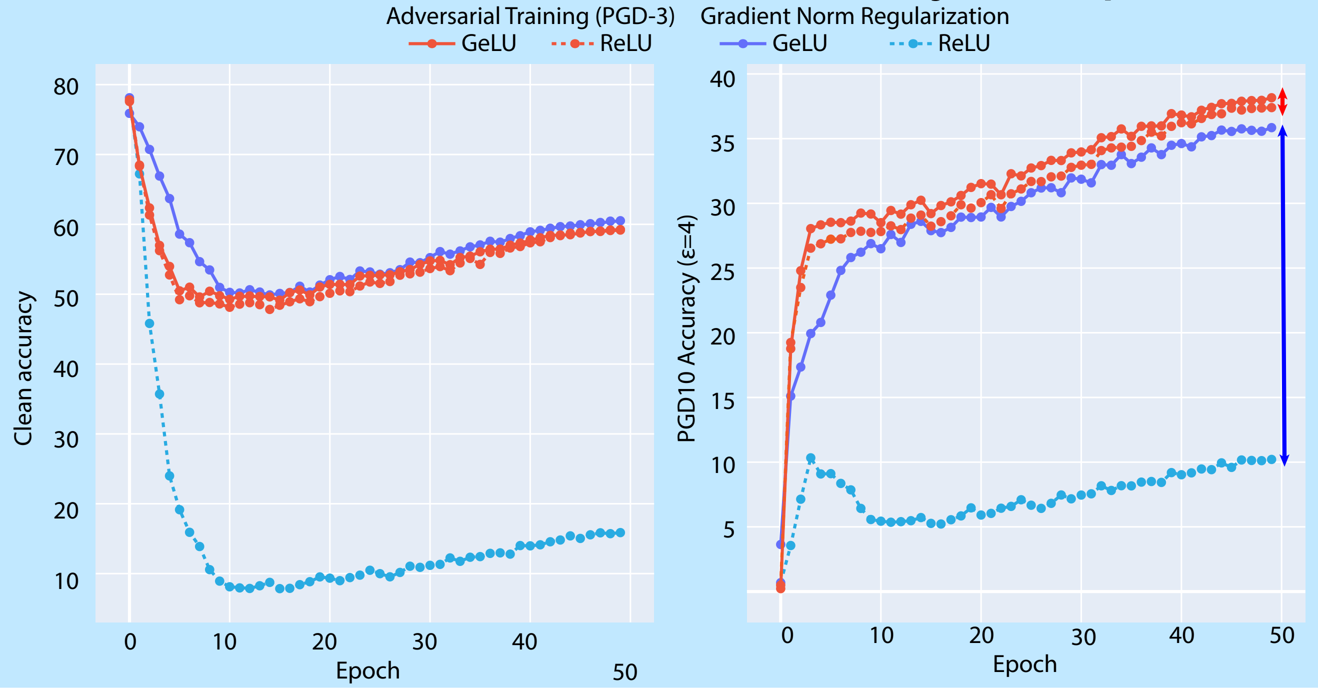
## Clean and AutoAttack accuracy

| Method | Clean | AutoAttack-$L_\infty$ | | |
|---|---|---|---|---|
| | - | $\epsilon = \frac{1}{255}$ | $\epsilon = \frac{2}{255}$ | $\epsilon = \frac{4}{255}$ |
| Natural Training | 84.19 | 00.00 | 00.00 | 00.00 |
| Grad. Norm ($\lambda_{\text{CE}} = 0.8, \lambda_{\text{GN}} = 1.2$) | 77.78 | 72.04 | 66.20 | 51.58 |
| Adv. Train. (PGD-3, $\epsilon = 4$) | 77.20 | 72.46 | 67.38 | 56.12 |

## PGD100 accuracy for $\epsilon \in [0, 32]$




## Clean and PGD10 accuracy vs epoch




## Loss-input gradient vs epoch