# Certification of Adversarial Robustness: Are We There Yet? What we Learn and Future Challenges

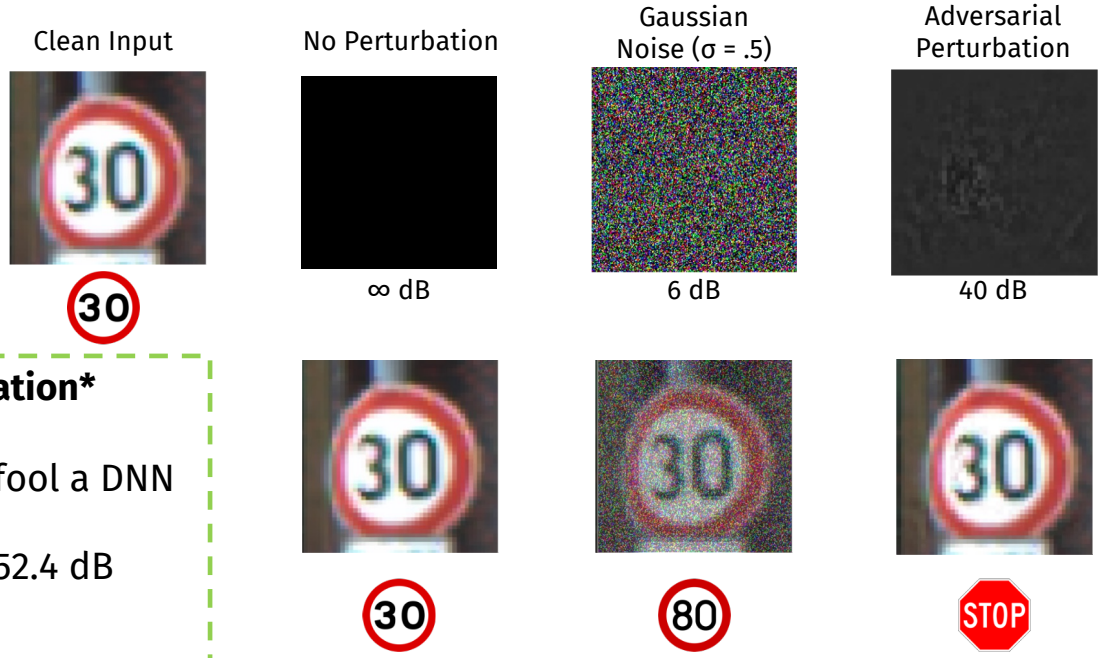
Workshop on AI Robustness and Safety for Cybersecurity

ELSA General Assembly 2025

**Fabio Brau - Assistant Professor**
University of Cagliari, Italy

# Adversarial Examples

A subtle, maliciously generated perturbation is sufficient to fool a classification model.




**Minimal Adversarial Perturbation***

Median perturbation sufficient to fool a DNN

**MNIST** : 25.9 dB, **CIFAR10**: 52.4 dB

**ImageNet**: 61.9 dB

* Deduced from [Jérôme Rony et al., Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses]

# Two Key Requirements for the Trustworthiness

**Improve Robustness to Attacks**

Reducing the sensitivity to adversarial examples improves the trustworthiness of a model.

**Certify the Adversarial Robustness**

Providing sound and reliable estimations of the robustness improves confidence in the behavior.

| Strategy | Improve Robustness | Certify Robustness | Adv-Train-Suited |
|---|---|---|---|
| Adversarial Attacks[1] | ✗ | ✓ (Prove Sensitivity) |  |
| Adversarial Training[2] | ✓ | ✗ | - |
| Verifications[3] | ✗ | ✓ (Prove Robustness) |  |
| Lipschitz Bounded[4] | ✓ | ✓ |  |
| Random. Smoothing[5] | ✓ | ✓ |  |

[1] Battista Biggio et al. Evasion Attacks against Machine Learning at Test Time (ECML PKDD 2013)
[2] Ian Goodfellow et al. Explaining and Harnessing Adversarial Examples (ICLR 2015)
[3] Guy Katz et al. *Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks (CAV 2017)*
[4] Moustapha Cisse et al. Parseval Networks: Improving Robustness to Adversarial Examples (ICML 2017)
[5] Jeremy Cohen et al. Certified adversarial robustness via randomized smoothing (NIPS 2019)

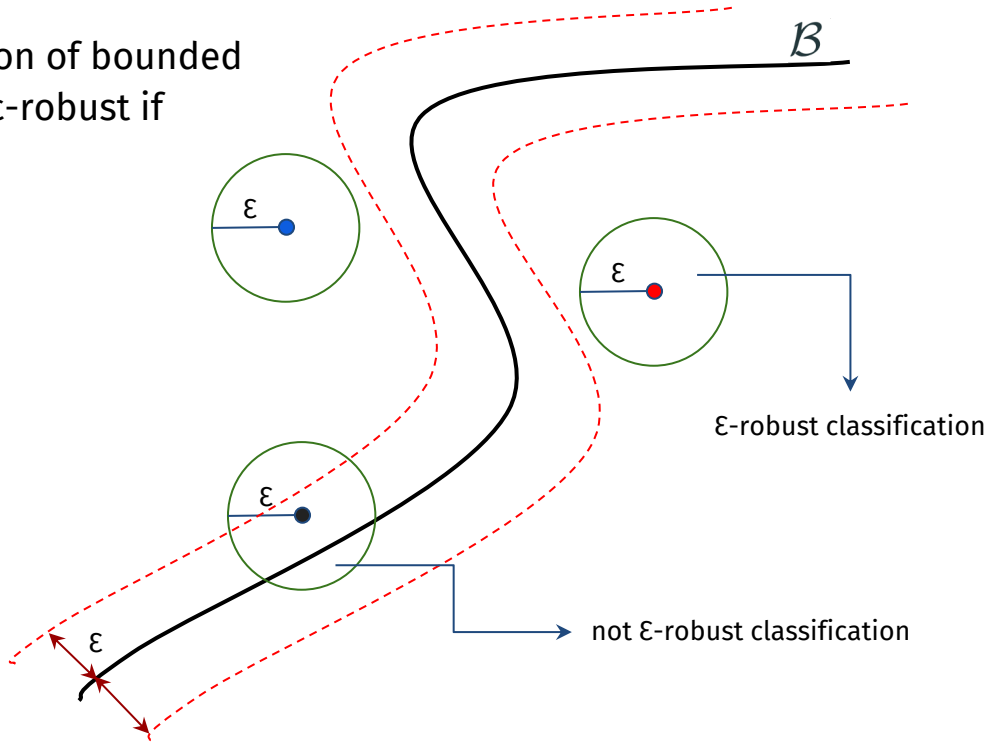
# The Certifiable $\varepsilon -$ Robust Accuracy

**Definition** (Robustness in $l^p$ norm)
Classification does not change under perturbation of bounded magnitude. In formulas, a classification $\mathcal{K}(x)$ is ε-robust if

$$\|\delta\| < \varepsilon \;\Rightarrow\; \mathcal{K}(x) = \mathcal{K}(x+\delta)$$

**Definition** (ε-robust accuracy)
Is the ratio of correct ε-robust classifications

$$\mathcal{A}_R(f,\varepsilon) = \mathbb{P}(\mathcal{K}_f(\mathbf{x}+\delta) = \mathcal{O}(\mathbf{x}),\ \forall\|\delta\| < \varepsilon)$$

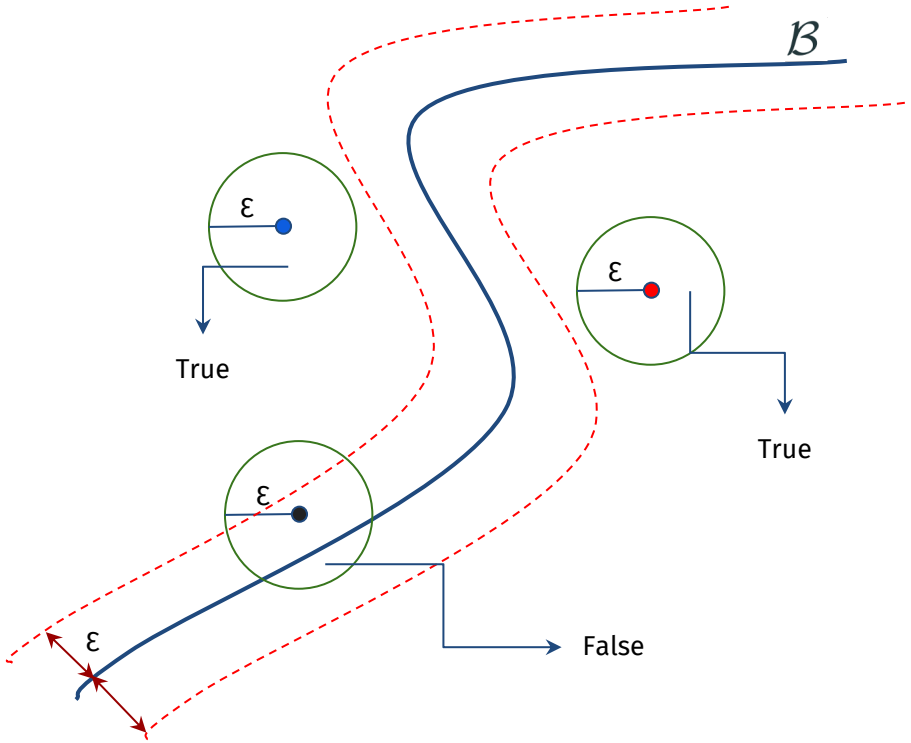


# The Verification of the Robustness

**Definition** (Verification of the robustness)
Given a classifier $\mathcal{K}$ and a sample $x$, check whether

$$\zeta(x): \quad \text{"}\forall y \in \mathcal{N}(x) \quad \mathcal{K}(x) = \mathcal{K}(y)\text{"}$$

where $\mathcal{N}$ is a neighborhood of $x$




**Theorem***
Let us assume $f$ a ReLU Deep Neural Network, and

$$\mathcal{N}(x) = \{y \in \mathbb{R}^n : \|y - x\|_\infty \leq \varepsilon\}$$

then completely check $\zeta(x)$ is **NP-HARD**

*Guy Katz et al. *Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks (CAV 2027)*

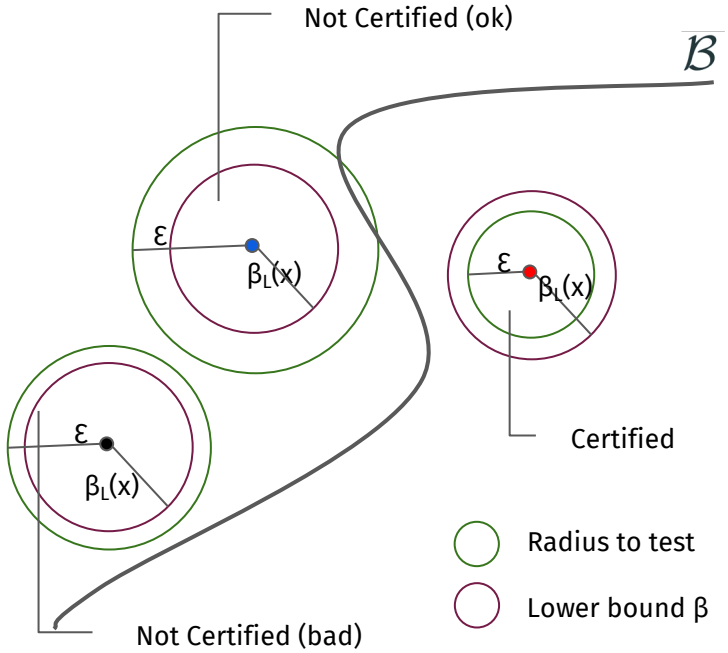
# Lipschitz-Based Certification

$$f(x_1) = \begin{bmatrix} 0.9 \\ 0.1 \\ 0 \end{bmatrix}$$

robust

$$f(x_2) = \begin{bmatrix} 0.6 \\ 0.4 \\ 0 \end{bmatrix}$$

unsafe




**Difference of Output-Logits and Robustness**

$$\varepsilon < \beta_L(x) := \frac{\min_{l \neq j} f_l(x) - f_j(x)}{L\sqrt{2}}$$

Qiyang Li et al. *Preventing Gradient Attenuation in Lipschitz Constrained Convolutional Networks*

# Certifiability by Design

Directly craft a model with a known Lipschitz Constant

**Observation** (Composition)
Composition of lipschitz functions is lipschitz

$$\underbrace{f(x) = f^{(k)} \circ f^{(k-1)} \circ \cdots \circ f^{(1)}(x)}_{L = \prod_{i=1}^{k} L_i}$$

**Examples of Lipschitz Layers**

Fully connected, Convolutional, Residual, Average and Maximum Pooling

**Remark 1**
The composition of 1-Lipschitz layers is 1-Lipschitz

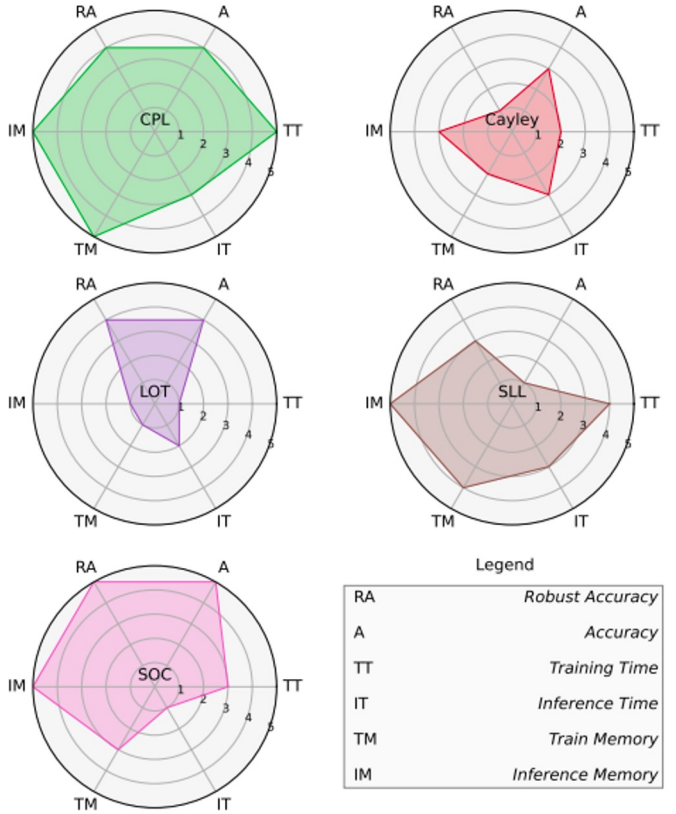
**Remark 2**
Common Perception Models are Lipschitz but with a very large constant

# CRA of Lipschitz-Bounded Models

**1-Lipschitz Layers Compared**

| Methods | Accuracy [%] XS | S | M | L | Robust Accuracy [%] XS | S | M | L |
|---|---|---|---|---|---|---|---|---|
| **CIFAR-10** | | | | | | | | |
| **AOL** | 71.7 | 73.6 | 73.4 | 73.7 | 59.1 | 60.8 | 61.0 | **61.5** |
| **BCOP** | 71.7 | 73.1 | 74.0 | 74.6 | 58.5 | 59.3 | 60.5 | **61.5** |
| **CPL** | 74.9 | 76.1 | 76.6 | 76.8 | 62.5 | 64.2 | 65.1 | **65.2** |
| **Cayley** | 73.1 | 74.2 | 74.4 | 73.6 | 59.5 | **61.1** | 61.0 | 60.1 |
| **LOT** | 75.5 | 76.6 | 72.0 | - | 63.4 | **64.6** | 58.7 | - |
| **SLL** | 73.7 | 74.2 | 75.3 | 74.3 | 61.0 | 62.0 | **62.8** | 62.3 |
| **SOC** | 74.1 | 75.0 | 76.9 | 76.9 | 61.3 | 62.9 | **66.3** | 65.4 |

The best CRA is **66%** for $\varepsilon = \frac{36}{255}$ (Naïve has $\approx 45\%$)



# Lipschitz-Bounded trained with Synthetic Data

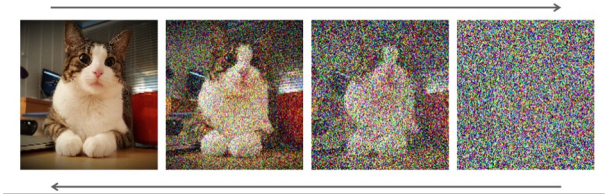
## Synthetic data improves the CRA

| | Clean Acc. (%) | VRA (%) at $\epsilon$ | | |
|---|---|---|---|---|
| | | $\frac{36}{255}$ | $\frac{72}{255}$ | $\frac{108}{255}$ |
| GloRo (Leino et al., 2021) | 77.0 | 58.4 | - | - |
| Local-Lip-B (+MaxMin) (Huang et al., 2021) | 77.4 | 60.7 | 39.0 | 20.4 |
| Cayley Large (Trockman & Kolter, 2021) | 74.6 | 61.4 | 46.4 | 32.1 |
| SOC 20 (Singla & Feizi, 2021) | 76.3 | 62.6 | 48.7 | 36.0 |
| CPL XL (Meunier et al., 2022) | 78.5 | 64.4 | 48.0 | 33.0 |
| AOL Large(Prach & Lampert, 2022) | 71.6 | 64.0 | 56.4 | 49.0 |
| SLL X-Large(Araujo et al., 2023) | 73.3 | 65.8 | 58.4 | 51.3 |
| GloRo LiResNet (+DDPM) (Hu et al., 2023) | 82.1 | 70.0 | - | - |
| GloRo CHORD LiResNet (+DDPM) | **87.0** | **78.1** | **66.6** | **53.5** |

CRA with no Synthetic Data

CRA with Synthetic Data

## Denoising Diffusion Models



50M DDPM-generated images (x $10^3$)

K. Hu et al. A Recipe for Improved Certifiable Robustness. (ICLR 2024)

# Future Challenges (?)

# Lipschitz-Levenshtein-Bounded DNNs for NLP

**Hamming Distance (same length sequences)**

**1-Lipschitz Models under the L-distance**

$$d_H(11001, 11000) = 1$$

Published as a conference paper at ICLR 2025

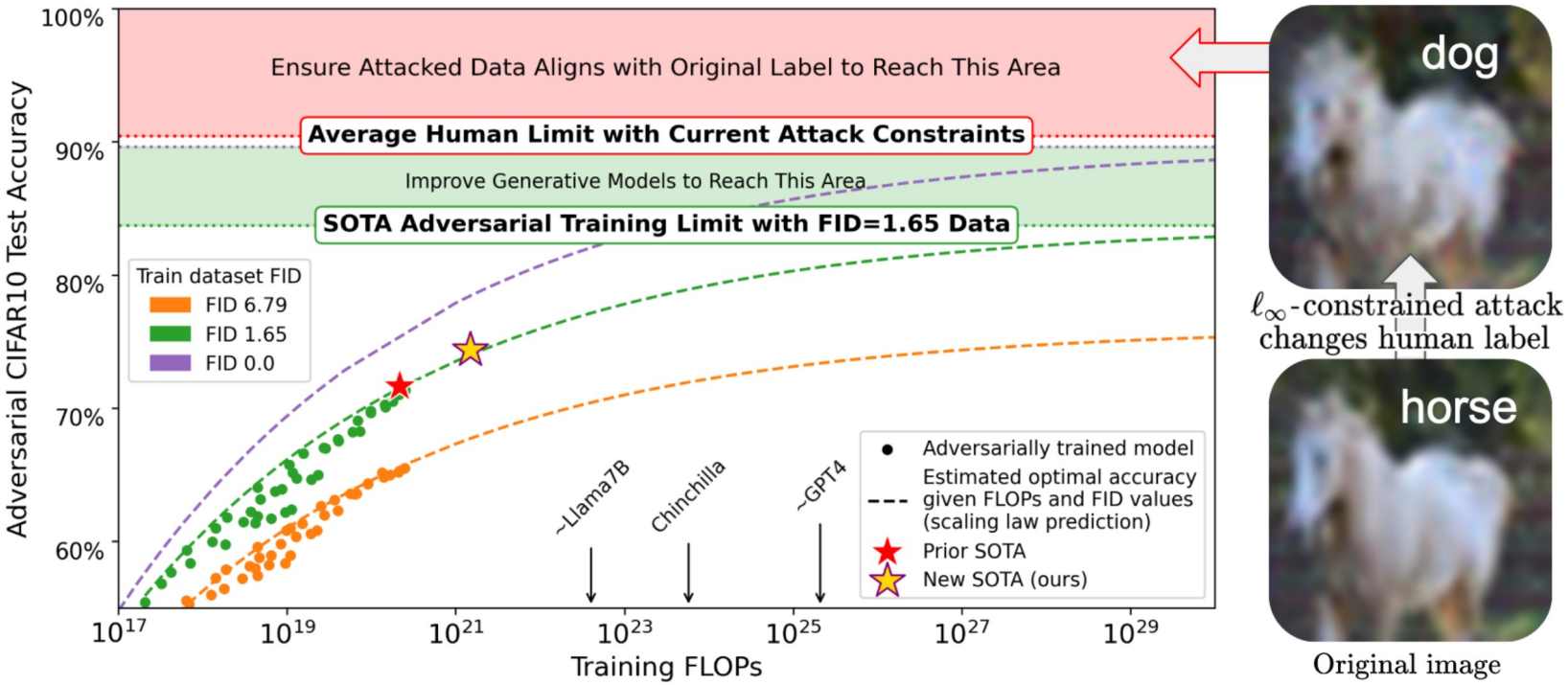
CERTIFIED ROBUSTNESS UNDER BOUNDED LEVENSHTEIN DISTANCE

Elias Abad Rocamora[EPFL], Grigorios G. Chrysos[W], Volkan Cevher[EPFL]

**Levenshtein Distance**

$$d_L(11001, 1100) = 1$$

# Scaling Law of Robustness: Is reaching h-CRA possible?




[7]BR Bartoldson et al. Adversarial robustness limits via scaling-law and human-alignment studies (ICML 2024)

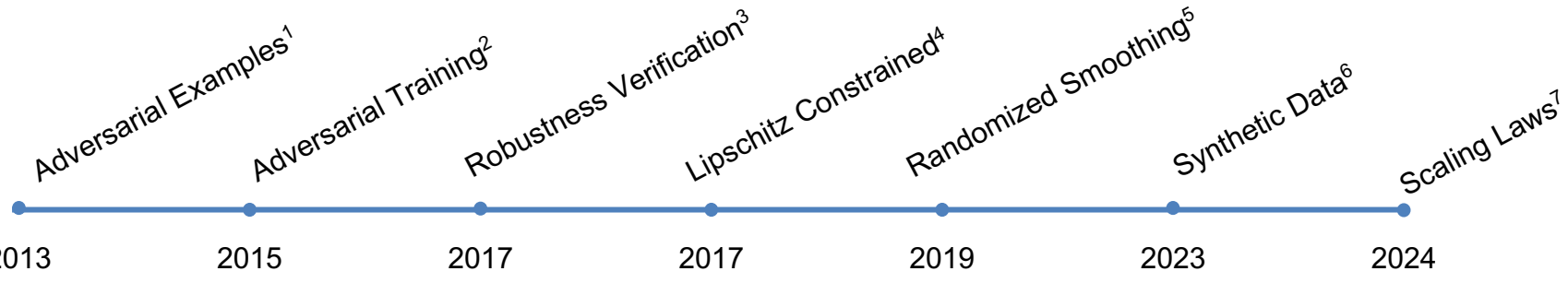
# Key Messages

Perturbations small $l_p$-norm may **exaggerate the threat**: real-world **robustness** may be **underestimated**.

Safety of critical systems may not be reachable with a stand-alone AI model

The Synthetic Data Generation plays a fundamental role in boosting performances.

# The Timeline of Adversarial Robustness




[1] Battista Biggio et al. Evasion Attacks against Machine Learning at Test Time (ECML PKDD 2013)
[2] Ian Goodfellow et al. Explaining and Harnessing Adversarial Examples (ICLR 2015)
[3] Guy Katz et al. *Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks (CAV 2017)*
[4] Moustapha Cisse et al. Parseval Networks: Improving Robustness to Adversarial Examples (ICML 2017)
[5] Jeremy Cohen et al. Certified adversarial robustness via randomized smoothing (NIPS 2019)
[6] Zekai Wang et al. Better Diffusion Models Further Improve Adversarial Training. (ICML 2023)
[7] BR Bartoldson et al. Adversarial robustness limits via scaling-law and human-alignment studies (ICML 2024)

sAIfer Lab

Joint lab on Safety and Security of AI

# Thanks for the Attention

Fabio Brau

University of Cagliari, Italy

UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI

Università di Genova