Defending Neural ODE Image Classifiers from Adversarial Attacks with Tolerance Randomization

Fabio Carrara[1], Roberto Caldelli[2,3], Fabrizio Falchi[1], Giuseppe Amato[1]
[1]ISTI CNR, Pisa, Italy — [2]CNIT, Florence, Italy — [3]Universitas Mercatorum, Rome, Italy

fabio.carrara@isti.cnr.it
https://github.com/fabiocarrara/neural-ode-features

MMForWild - ICPR 2021 - January, 11th
Milan, Italy (Virtual)

Universitas Mercatorum
Università telematica delle Camere di Commercio Italiane

AIMH Lab
Artificial Intelligence for Media and Humanities

Summary

- Neural ODEs
  - what they are
  - how can be used
  - why they are interesting (adaptivity and the tolerance parameter)

- Carlini & Wagner Adversarial Attack
  - the gist of it
  - how Neural ODEs respond

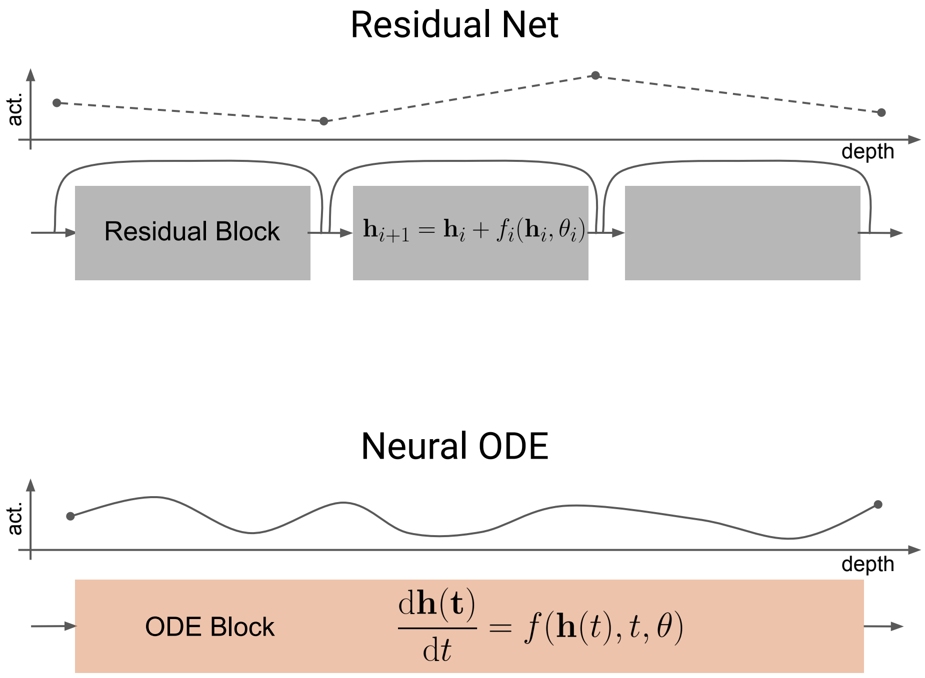
- Tolerance Randomization
  - an adversarial detection scheme for Neural ODEs under strong adversarials inputs
  - experiments and results

- Conclusions and Future Work

Neural Ordinary Differential Equations [9]

- Generalization of Residual Networks
  - **ResNet**: discrete number of coarse updates
  - **N-ODE**: continuous and smooth evolution (infinitesimal updates) defined by parametric ODE

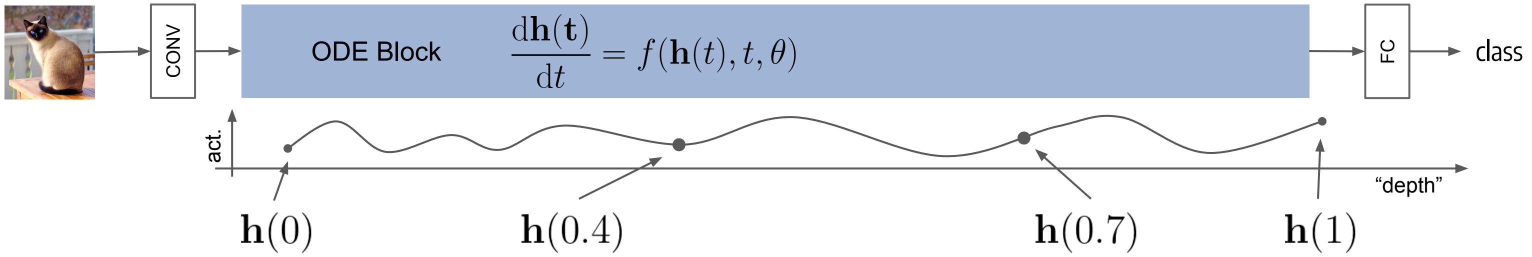
- **Forward**: solve with ODE solver
- **Output**: final step of the solution
- **Fully Differentiable**: train the params of ODE with SGD



[9] Chen, Tian Qi, et al. "Neural ordinary differential equations." Advances in neural information processing systems. 2018.

Neural ODE Image Classifiers

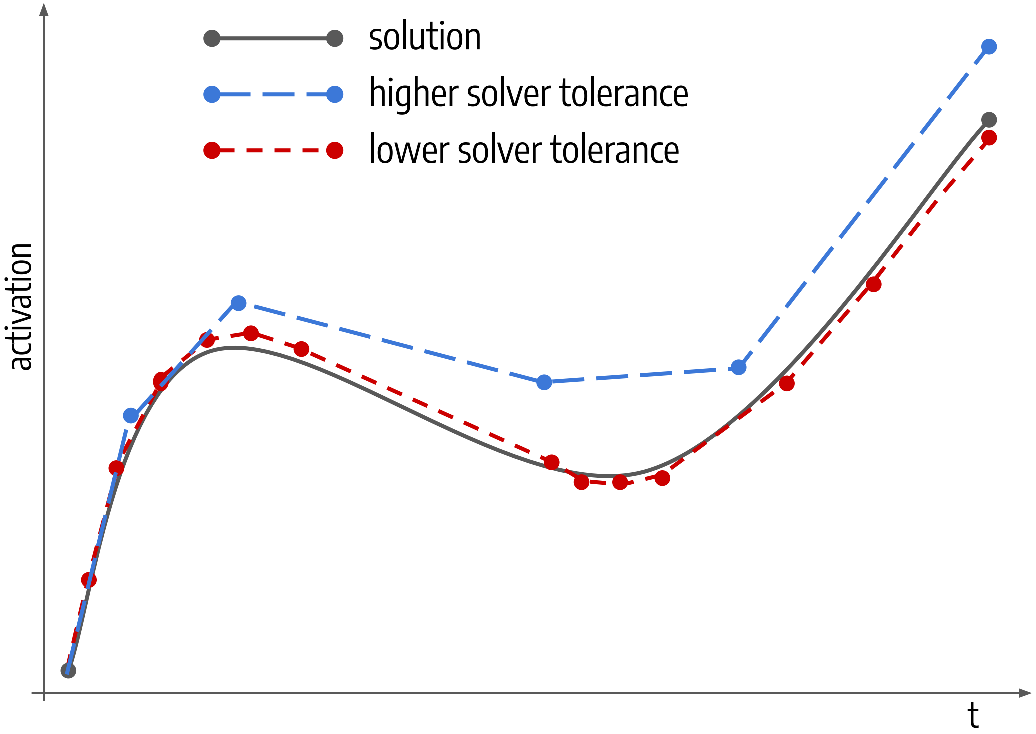
- Neural ODE for Image Classification



- $f(\mathbf{h}(t), t, \theta)$ is implemented as **a small convnet** (comparable to a residual block)
- in the **forward pass**, an **ODE solver is used** to find the output $\mathbf{h}(1)$
- in the training phase, we **learn dynamics** (by optimizing θ with SGD) that evolve inputs to discriminative features **for classification**
- performance comparable to standard convnet models

Neural ODE Adaptivity

- ODE Solvers
  - compute solution by taking small steps in time
- Adaptive ODE Solvers
  - step size is adaptively chosen at each iteration
- Tolerance parameter τ
  - controls the **speed-precision trade-off** of the solver
  - ***high*** $\boldsymbol{\tau}$ $\Rightarrow$ less steps, less precise & less computational expensive solution
  - ***lower*** $\boldsymbol{\tau}$ $\Rightarrow$ more steps, more precise solution, more compute needed

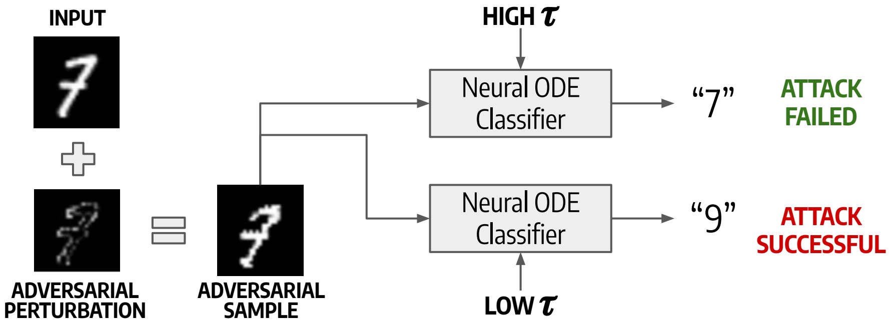


Effects of Tolerance

- Tolerance τ affects classification performance
  - MNIST and CIFAR-10
  - ResNet as benchmark
  - $\tau_{train} = 10^{-3}$, τ_{test} varies
  - **Classification Error vs τ_{test}**

	ResNet	Neural ODE (τ)				
MNIST		10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
Classification Error (%)	0.4	0.5	0.5	0.6	0.8	1.2
CIFAR-10						
Classification Error (%)	7.3	9.1	9.2	9.3	10.6	11.3

- Tolerance τ affects adversarial robustness [5]
  - high $\tau \Rightarrow$ robustness increases vs weak attacks (PGD)
  - adversarial perturbation is more difficulty propagated through the network



[5] Carrara, F., Caldelli, R., Falchi, F. and Amato, G., 2019, December. *On the robustness to adversarial examples of neural ode image classifiers*. In 2019 IEEE International Workshop on Information Forensics and Security (**WIFS '19**) (pp. 1-6). IEEE.

Carlini and Wagner (CW) Attack

- Proposed by Carlini and Wagner [3]
  - Considered a strong attack
  - bypassed several proposed defenses for standard neural networks

- Optimization-based attack
  - $\mathbf{x}$ is the natural sample
  - $\mathbf{x}^{\text{adv}}$ is the adversarial sample
  - $g()$ is the misclassification objective
  - $|| \mathbf{x}^{\text{adv}} - \mathbf{x} ||_2$ is the magnitude of the perturbation
  - c is grid-searched

$$\min \left(c \cdot \underbrace{g\left(\mathbf{x}^{\text{adv}}\right)}_{\textbf{misclassification objective}} + \overbrace{\left\|\mathbf{x}^{\text{adv}} - \mathbf{x}\right\|_2^2}^{\textbf{small perturbation objective}} \right)$$

- Usually finds very small perturbations leading to misclassification

[3] Carlini, N., Wagner, D., Towards evaluating the robustness of neural networks. In 2017 IEEE SP. pp. 39-57, 2017

Neural ODE vs CW Attacks

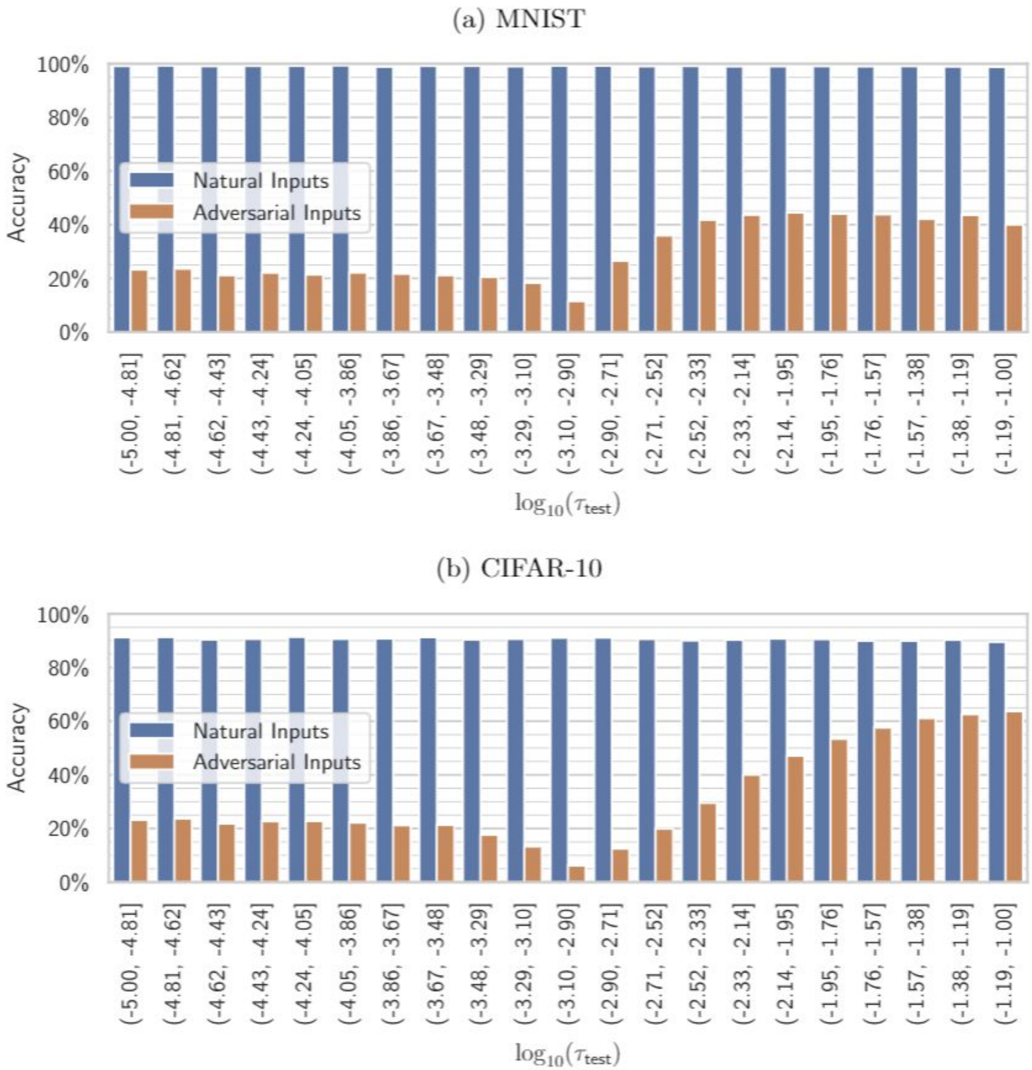
- Neural ODEs are still vulnerable
  - MNIST and CIFAR-10
  - Carlini and Wagner (CW) Adversarial Attack
  - $\tau_{attack} = \tau_{test}$
- How τ affects robustness to CW attacks?
  - Attack Success Rate vs τ
  - Mean Adversarial Perturbation Norm vs τ
  - **higher τ ⇒**
    - **lower attack success rate**, or
    - **higher perturbation magnitude**

	ResNet	Neural ODE (τ)				
MNIST		10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
Classification Error (%)	0.4	0.5	0.5	0.6	0.8	1.2
Attack Success Rate (%)	99.7	99.7	90.7	74.4	71.6	69.7
Mean L2 Perturb (x10^{-2})	1.1	1.4	1.7	1.9	1.7	1.9
CIFAR-10						
Classification Error (%)	7.3	9.1	9.2	9.3	10.6	11.3
Attack Success Rate (%)	100	100	100	100	100	100
Mean L2 Perturb (x10^{-5})	2.6	2.2	2.4	4.1	8	13.7

Attacking & Defending

- **Attack assumption**:
  assuming no defense, the best strategy for an attacker is to set
  $\boldsymbol{\tau}_{\textbf{attack}} = \boldsymbol{\tau}_{\textbf{train}}$

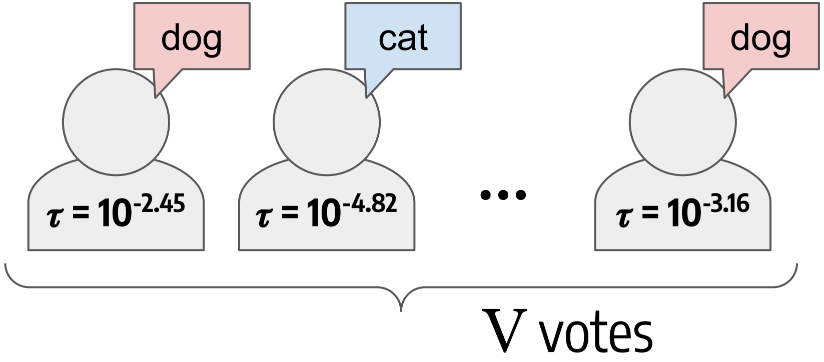
- **Defense strategy**:
  use $\boldsymbol{\tau}_{\textbf{test}} \neq \boldsymbol{\tau}_{\textbf{train}}$ in prediction
  - increased robustness
  - negligible performance drop

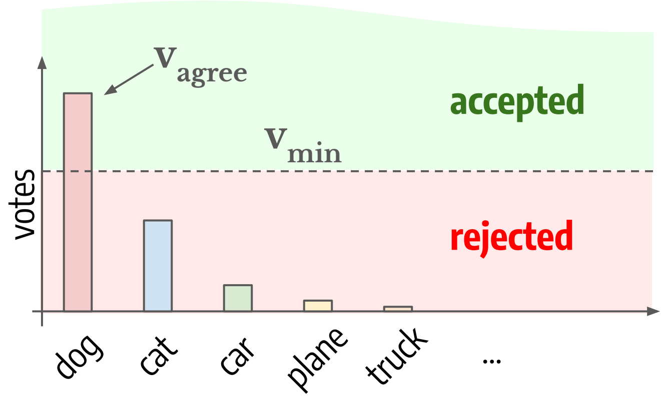


Tolerance Randomization Defense

- Randomize $\boldsymbol{\tau}_{\textbf{test}}$ at prediction time
  - Randomly sample $\boldsymbol{\tau}_{\textbf{test}}$ from log-uniform interval $[10^{-5}; 10^{-1}]$
  - Perform the prediction V times on the same image with the same model using the sampled $\boldsymbol{\tau}_{\textbf{test}}$

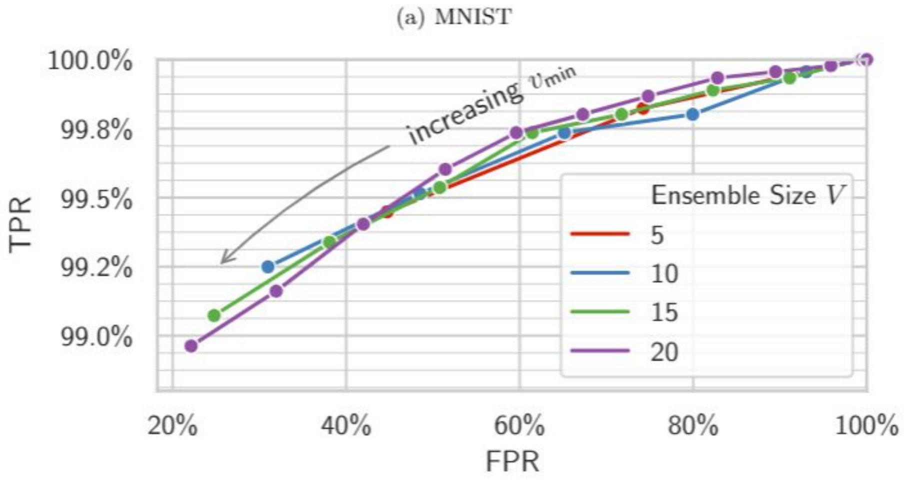
- Create an Ensemble
  - **Super-Majority (Qualified Majority) Voting**
  - $\mathbf{v}_{\mathbf{agree}}$ = number of votes given to the most voted class
  - $\mathbf{v}_{\mathbf{min}}$ = minimum number of votes needed to accept a class
  - if $\mathbf{v}_{\mathbf{agree}} > \mathbf{v}_{\mathbf{min}}$, we accept the classification, otherwise we discard it (may be adversarial)

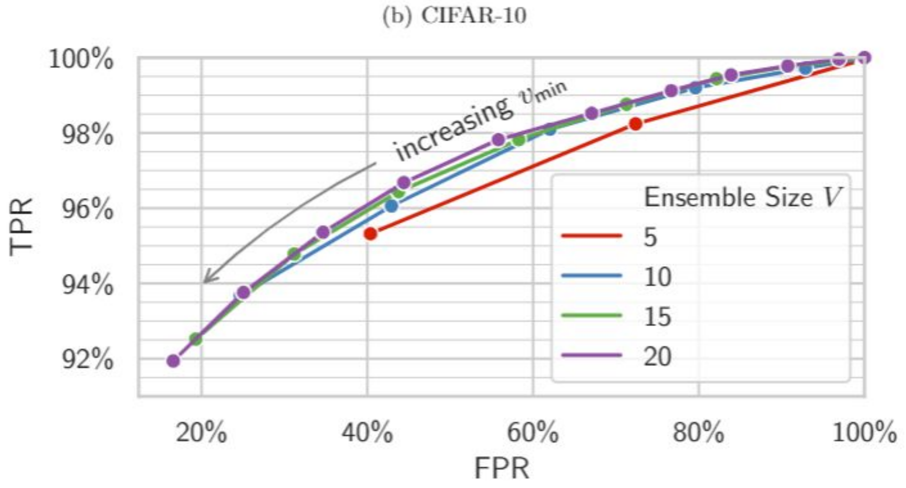




Experiment and Results

- MNIST and CIFAR-10
- Neural ODE Image Classifier ($\tau_{train} = 10^{-3}$)
- Carlini and Wagner attacks ($\tau_{attack} = \tau_{train}$)
    - 5.000 pristine + 5.000 adversarial images
- Tolerance τ_{test} randomized in log-uniform interval $[\,10^{-5}\,;\,10^{-1}\,]$
- Ensemble Size V = {5, 10, 15, 20}
- ROC Curve varying v_{min}
    - Positive = Natural
    - Negative = Adversarial





Conclusions and Future Work

- Analysis of Neural ODE as image classifiers robust to adversarial example
- We proposed **Tolerance Randomization** for defending Neural ODEs
- Preliminary experiments on white-box, zero-knowledge attacks:
    - reject ∼80% strong Carlini and Wagner adversarials images
    - accept +90% pristine images
- **Future Work**
    - thorough analysis of attack robustness under (τ_{train}, τ_{test}, τ_{attack}) decoupling
    - devise attacks and defenses for more stringent scenarios (attacker knows about defense)

fabio.carrara@isti.cnr.it https://github.com/fabiocarrara/neural-ode-features