Accurate, reliable and fast robustness evaluation

Wieland Brendel[1–3] • Jonas Rauber[1–3] • Matthias Kümmerer[1–3] • Ivan Ustyuzhaninov[1–3] • Matthias Bethge[1–5]

[1] Centre for Integrative Neuroscience, University of Tübingen, Germany • [2] Bernstein Center for Computational Neuroscience, University of Tübingen, Germany • [3] Institute for Theoretical Physics, University of Tübingen, Germany • [4] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[5] Center for Neuroscience Artificial Intelligence, Baylor College of Medicine, Houston, USA

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

TL;DR

PROBLEM
Adversarial attacks often overestimate robustness of ML models because of optimisation issues.

SOLUTION
Novel attack that follows decision boundary and solves inner trust-region optimisation problem to find optimal step.

BENEFITS?
Finds smaller adversarials in less steps than SOTA on L0, L1, L2 & Linf with almost no hyperparameter tuning. More robust to gradient masking.

CODE?
Use it soon with Foolbox.

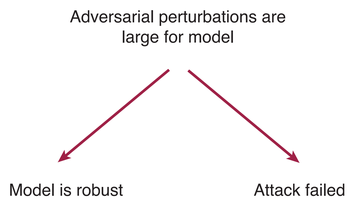
Overview

Progress towards more adversarially robust models is significantly impaired by the difficulty of evaluating the robustness of ML models. Today's methods are either fast but brittle (gradient-based attacks), or they are fairly reliable but slow (score- and decision-based attacks). We here develop a new set of gradient-based adversarial attacks for L0, L1, L2 and Linf which

(a) are more reliable in the face of gradient-masking than other gradient-based attacks,
(b) perform better and are more query efficient than current state-of-the-art gradient-based attacks,
(c) can be flexibly adapted to a wide range of adversarial criteria and
(d) require virtually no hyperparameter tuning.

Implementations will soon be available in Foolbox, CleverHans & ART.

The devil of model robustness



Optimal step within trust-region

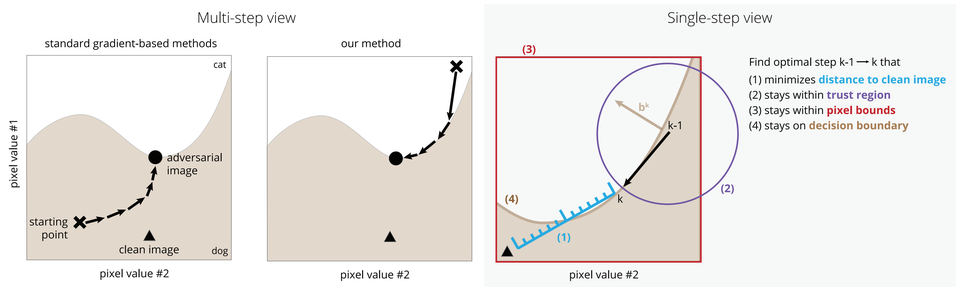
$$\min_{\boldsymbol{\delta}} \left\| \boldsymbol{x} - \tilde{\boldsymbol{x}}^{k-1} - \boldsymbol{\delta}^k \right\|_p$$ minimize distance

s.t. $0 \leq \tilde{\boldsymbol{x}}^{k-1} + \boldsymbol{\delta}^k \leq 1$ stay within bounds

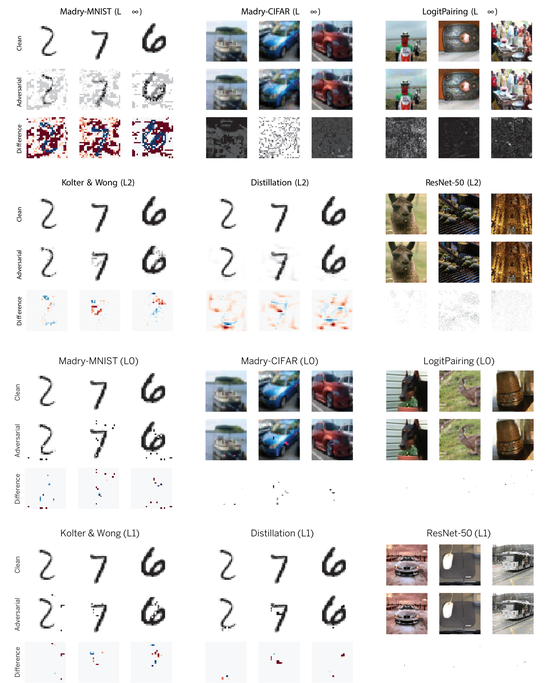
$\boldsymbol{b}^{k\top} \boldsymbol{\delta}^k = c^k$ move to boundary

$\left\| \boldsymbol{\delta}^k \right\|_2^2 \leq r$ stay within trust-region

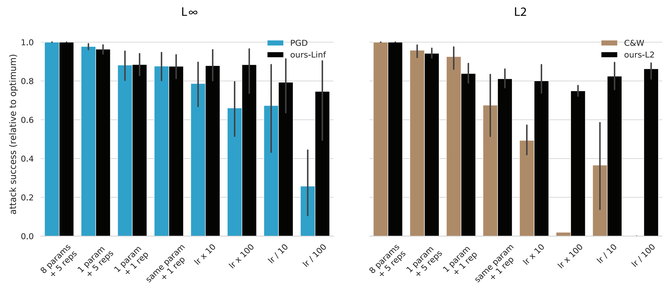
Our attack moves along the decision boundary



Compared to SOTA, our attack finds better minima in less queries



Attack needs almost no hyperparameter tuning



Samples of our adversarial attack

Algorithm

Algorithm 1: Overview over the trust-region solver for a given L_p norm.

Data: clean image $\boldsymbol{x}$, perturbed image $\tilde{\boldsymbol{x}}$, boundary $\boldsymbol{b}$, logit-difference c, trust region r
Result: optimal perturbation $\boldsymbol{\delta}$ minimizing (1)
begin
  $\mu_0, \lambda_0 \longleftarrow 0, 0$
  while *not converged* **do**
    $g(\lambda_k, \mu_k) \longleftarrow \inf_{\boldsymbol{\delta}} \Lambda(\boldsymbol{\delta}, \mu_k, \lambda_k) \quad \text{s.t.} \quad u \leq \tilde{\boldsymbol{x}} + \boldsymbol{\delta} \leq \ell$
    $\nabla g(\lambda_k, \mu_k) \longleftarrow \nabla \inf_{\boldsymbol{\delta}} \Lambda(\boldsymbol{\delta}, \mu_k, \lambda_k) \quad \text{s.t.} \quad u \leq \tilde{\boldsymbol{x}} + \boldsymbol{\delta} \leq \ell$
    $\mu_{k+1}, \lambda_{k+1} \longleftarrow \text{BFGS-B}(g(\lambda_k, \mu_k), \nabla g(\lambda_k, \mu_k))$
  end
  $\boldsymbol{\delta}^* \leftarrow \operatorname{arginf}_{\boldsymbol{\delta}} \Lambda(\boldsymbol{\delta}, \mu_k, \lambda_k) \quad \text{s.t.} \quad u \leq \tilde{\boldsymbol{x}} + \boldsymbol{\delta} \leq \ell$
end

Conclusions

- Unlike other attacks, our methods follows the decision boundary to find optimal adversarial perturbations.
- Compared to SOTA, our attack finds smaller adversarial perturbations across a wide range of models in several Lp-metrics.
- Our attack is particularly well suited for adversarially trained models as it moves along the area where maximal signal in the gradients can be expected.

Code & Links

code (github) paper (arxiv)

Funding

- Tübingen AI Center (FKZ: 01IS18039A)
- German Research Foundation (DFG): Collaborative Research Center (CRC 1233) "Robust Vision"
- Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center

CIN Werner Reichardt Centre for Integrative Neuroscience

bccn tübingen

Contact & Social Media

wieland.brendel@uni-tuebingen.de

@wielandbr @wielandbr