Accurate, reliable and fast robustness evaluation

Overview

Progress towards more adversarially robust models is significantly impeded 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 \( L_0, L_1, L_2 \) and \( \infty \) which
- are more reliable in the face of gradient-masking than other gradient-based attacks,
- perform better and are more query efficient than current state-of-the-art gradient-based attacks,
- can be flexibly adapted to a wide range of adversarial criteria and
- require virtually no hyperparameter tuning.

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

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 \( L_0, L_1, L_2 \) & \( \infty \) with almost no hyperparameter tuning. More robust to gradient masking.

CODE?
Use it soon with Foolbox.

TL;DR
- The devil of model robustness
- Adversarial perturbations are large for model
- Model is robust: Attack failed
- Optimal step within trust-region

Algorithm

Find optimal step \( \theta \) that
1. Minimizes distance to clean image
2. Stays within trust region
3. Stays within pixel bounds
4. Stays on decision boundary

Our attack moves along the decision boundary

Attack needs almost no hyperparameter tuning

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

Conclusion

- 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.

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

Contact & Social Media
wieland.brendel@uni-tuebingen.de
code (github) paper (arxiv)
@wielandbr

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
- Center for Neuroscience Artificial Intelligence, Baylor College of Medicine, Houston, USA

The devil of model robustness

Adversarial perturbations are large for model

Model is robust: Attack failed

Optimal step within trust-region

\[
\min_{\theta} \| \theta - \theta^{k-1} \|_p \quad \text{minimizes distance}
\]
\[
\text{s.t.} \begin{align*}
0 &\leq \theta - 1 \leq 1, \\
\theta^* &\neq \theta, \\
\| \theta^* \|_p &\leq \tau \\
\| \theta - \theta^{k-1} \|_p &\leq \epsilon 
\end{align*}
\]