Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples

Maura Pintor, Luca Demetrio, Angelo Sotgiu, Giovanni Manca, Ambra Demontis, Nicholas Carlini, Battista Biggio, Fabio Roli

"A Blessing in Disguise: The Prospects and Perils of Adversarial Machine Learning" ICML workshop 2021

Pattern Recognition and Application Laboratory

Department of Electrical and Electronic Engineering
University of Cagliari, Italy

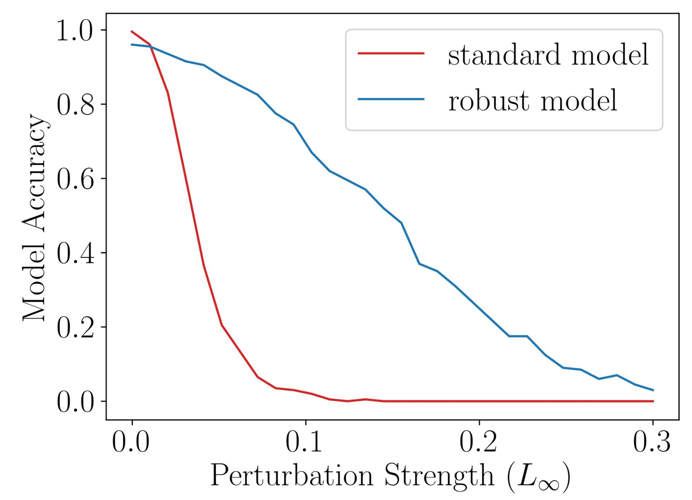
Pluribus One
seeing one in many

Google

Adversarial Robustness Evaluation

Goal:
- ➔ Find adversarial examples with a given perturbation budget
- ➔ Evaluate the robust accuracy

Problems:
- ➔ We have to rely on empirical evaluations
- ➔ Attacks often fail
- ➔ *False sense of security*
- ➔ Hard to fix: only guidelines but no practical debugging tools available!



Gradient-based Attacks

- ➔ General formalization for untargeted and targeted attacks
- ➔ We highlight steps related to common failures

Input : $\boldsymbol{x}$, the initial point; y_t, the target (true) class label if the attack is targeted (untargeted); n, the number of iterations; α, the learning rate; f, the target model; $(\boldsymbol{x}_{lb}, \boldsymbol{x}_{ub})$, the bounds of the input space; Δ, the considered region.

Output : $\boldsymbol{x}^\star$, the solution found by the algorithm

1 $\boxed{\boldsymbol{x}_0 \leftarrow init(\boldsymbol{x})}$ ▷ `Initialize starting point`

2 $\boxed{\hat{\boldsymbol{\theta}} \leftarrow approximation(\boldsymbol{\theta})}$ ▷ `Approximate model's parameters`

3 $\boldsymbol{\delta}_0 \leftarrow \mathbf{0}$ ▷ `Initial` δ

4 $\boxed{\textbf{for } i \in [1, n] \textbf{ do}}$

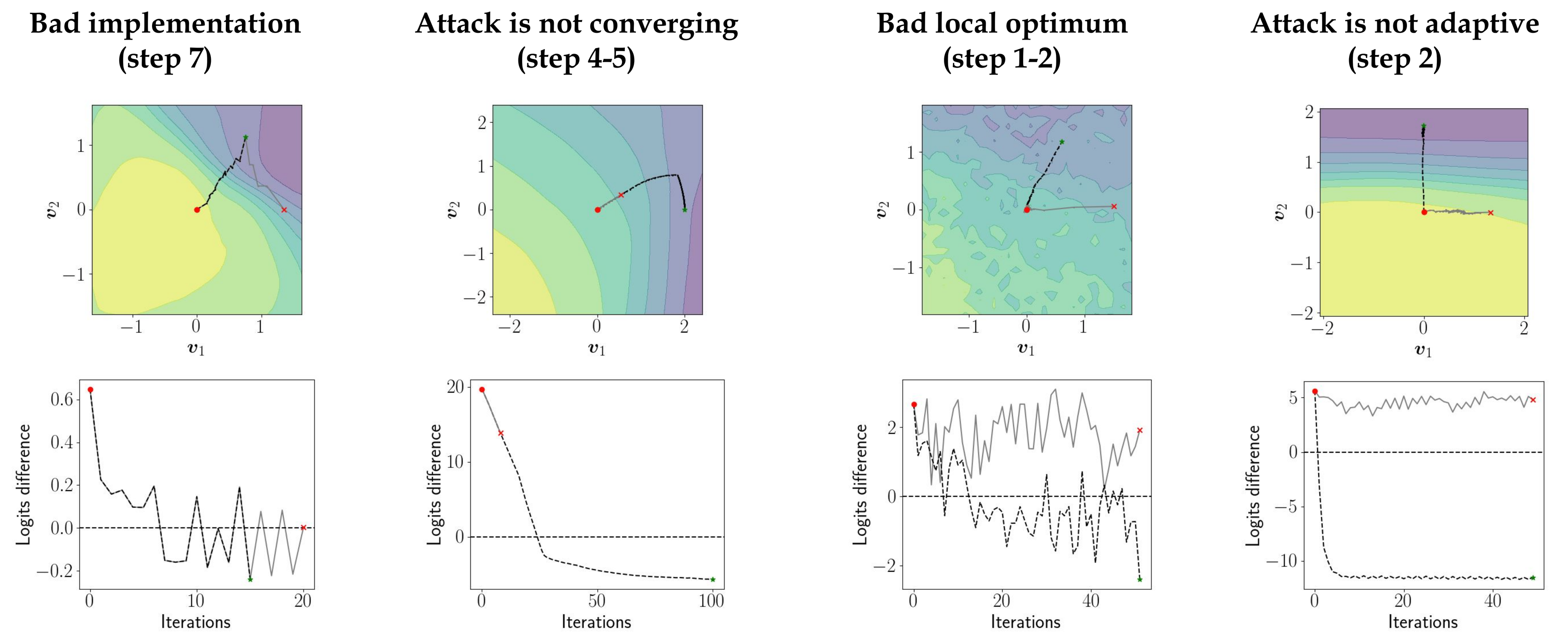
5 $\quad \boxed{\boldsymbol{\delta}' \leftarrow \boldsymbol{\delta}_i + \alpha \nabla_{\boldsymbol{x}_i} L(\boldsymbol{x}_0 + \boldsymbol{\delta}_i, y_t; \hat{\boldsymbol{\theta}})}$ ▷ `Compute optimizer step`

6 $\quad \boldsymbol{\delta}_{i+1} \leftarrow apply\_constraints(\boldsymbol{x}_0, \boldsymbol{\delta}', \Delta, \boldsymbol{x}_{lb}, \boldsymbol{x}_{ub})$ ▷ `Apply constraints`

7 $\boxed{\boldsymbol{\delta}^\star \leftarrow best(\boldsymbol{\delta}_0, ..., \boldsymbol{\delta}_n)}$ ▷ `Choose best perturbation`

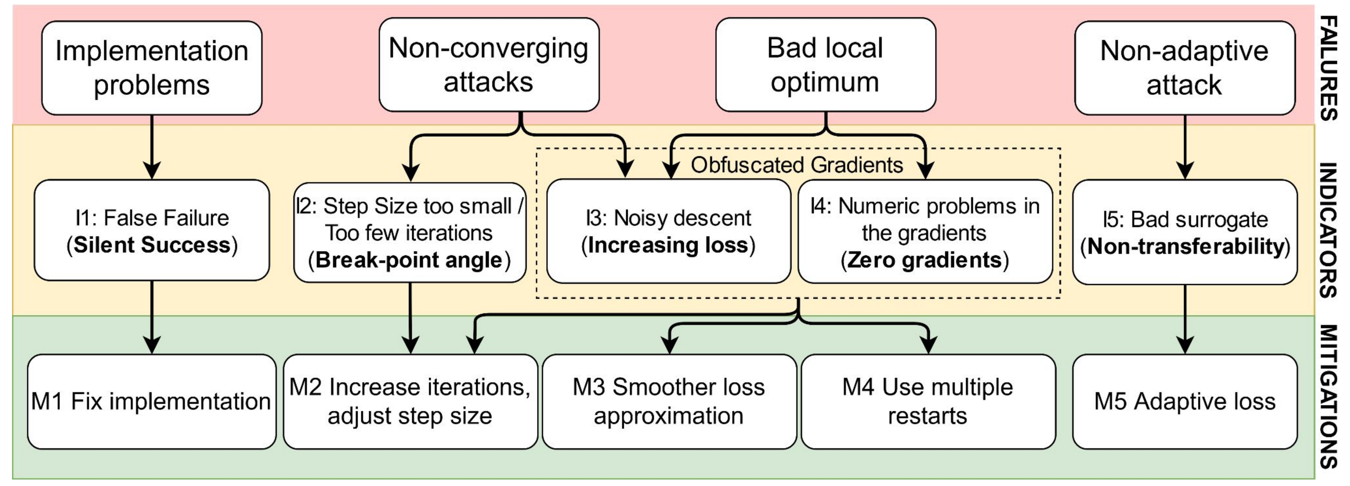
8 **return** $\boldsymbol{\delta}^\star$

Attack failures



Indicators and mitigations

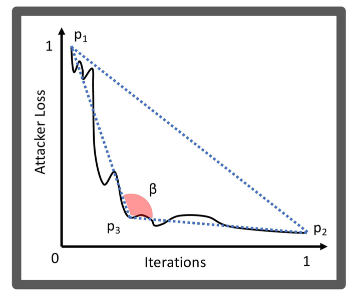
- ➔ We formulate 5 quantitative indicators (all in [0, 1])
- ➔ Each indicator is related to one or more failure
- ➔ We also propose 5 mitigations to apply, based on indicators results



Break point angle

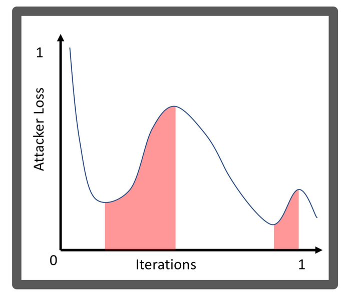
$1 - |\cos \beta|$

the attack loss is normalized



Increasing loss

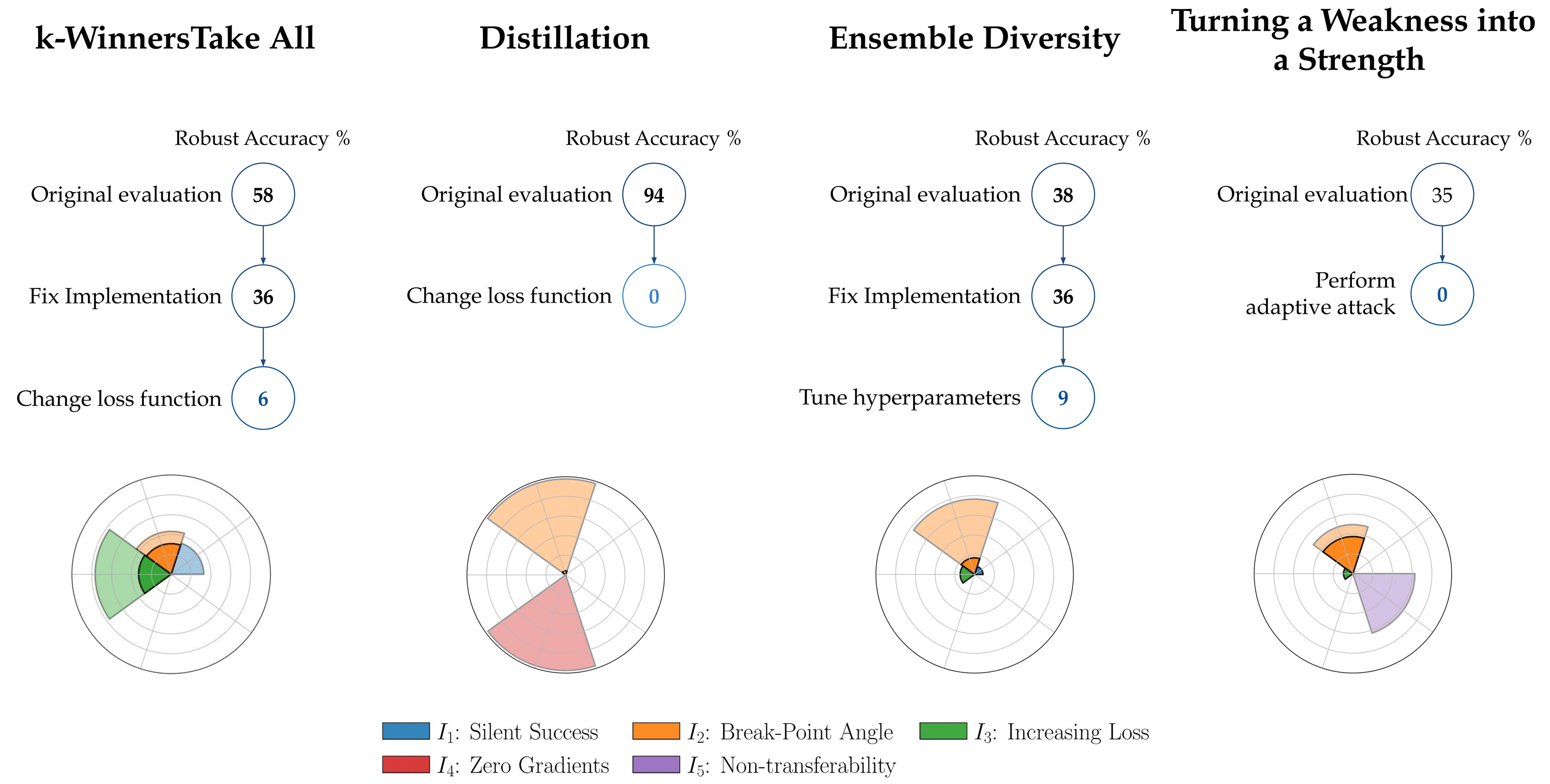
Area under the attack loss if increasing

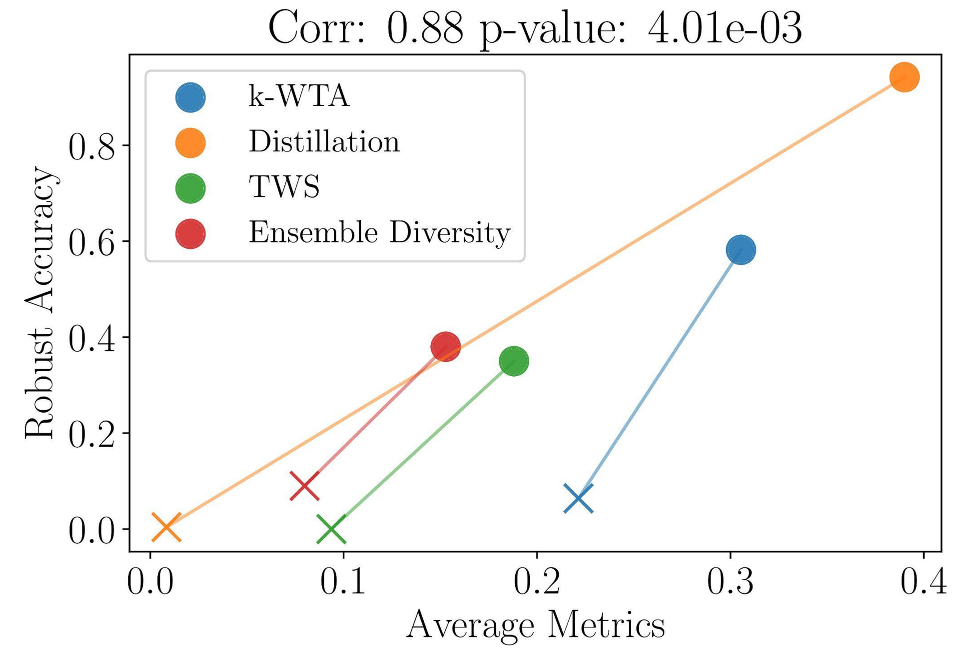


Experiments

Setting:
- ➔ We select 4 defenses with reported failures
- ➔ We evaluate our indicators
- ➔ We apply mitigations

Results:
- ➔ Indicators correctly reveal the "false sense of security"
- ➔ Patched attacks drop robust accuracy
- ➔ Indicators are strongly correlated with attacks performance





Useful links and implementations

- ➔ Open source code https://github.com/pralab/IndicatorsOfAttackFailure
- ➔ Paper available https://arxiv.org/abs/2106.09947
- ➔ Implemented with SecML

SECML
https://secml.gitlab.io/
Twitter: @secml_py

Key Takeaways

- ➔ Unified framework for gradient-based attacks and categorization of main failures
- ➔ Framework for debugging faulty-conducted security evaluations with quantitative indicators and mitigations strategies
- ➔ Empirical evaluation on 4 case-studies
  - ◆ indicators highlight failures
  - ◆ mitigations improve the robustness evaluation

Future Work

- ➔ Integration in benchmarks
- ➔ Add more indicators
- ➔ Further automatization
  - ◆ Towards MLSecOps

References

[1] Tramèr, F., Carlini, N., Brendel, W., & Madry, A. (2020). **On Adaptive Attacks to Adversarial Example Defenses.** *ArXiv, abs/2002.08347.*
[2] Xiao, C., Zhong, P., & Zheng, C. (2019). **Resisting Adversarial Attacks by k-Winners-Take-All.** *ArXiv, abs/1905.10510.*
[3] Papernot, N., Mcdaniel, P., Wu, X., Jha, S., & Swami, A. (2016). **Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks.** *2016 IEEE Symposium on Security and Privacy (SP), 582-597.*
[4] Pang, T., Xu, K., Du, C., Chen, N., & Zhu, J. (2019). **Improving Adversarial Robustness via Promoting Ensemble Diversity.** *ArXiv, abs/1901.08846.*
[5] Yu, T., Hu, S., Guo, C., Chao, W., & Weinberger, K.Q. (2019). **A New Defense Against Adversarial Images: Turning a Weakness into a Strength.** *ArXiv, abs/1910.07629.*
[6] Melis, M., Demontis, A., Pintor, M., Sotgiu, A., & Biggio, B. (2019). **secml: A Python Library for Secure and Explainable Machine Learning.** *ArXiv, abs/1912.10013.*

Department of Electrical and Electronic Engineering
University of Cagliari

Piazza d'Armi - 09123 Cagliari, Italy
https://pralab.diee.unica.it