ICML
International Conference
On Machine Learning

SORA: Free Second-Order Attacks in Fast Adversarial Training

International Conference on Machine Learning 2026

Mazdak Teymourian*, Ramtin Moslemi*,
Farzan Rahmani, Mohammad Hossein Rohban

Robust and Interpretable Machine Learning Lab @ Sharif University of Technology

Seoul, South Korea | July 2026

Adversarial Training

- Mądry et al.[1] formulate the min–max optimization problem:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{\boldsymbol{\delta}\in\Delta} \mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{\delta}), y) \right]$$

- Mądry et al.[1] use the **expensive** multi-step attack PGD.
- We can use the **Fast** Gradient Sign Method (FGSM)[2] instead:

$$\boldsymbol{\delta} = \epsilon \cdot \operatorname{sign}\left(\nabla_{\boldsymbol{x}} \mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)\right),$$

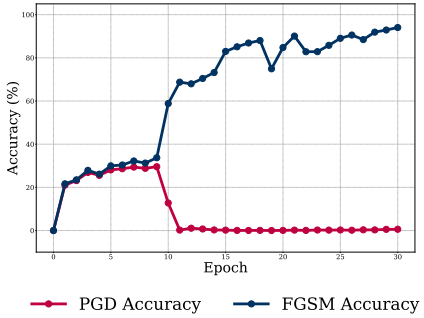
but there is a problem ...

[1] Mądry et al. Towards Deep Learning Models Resistant to Adversarial Attacks. **ICLR**, 2018.
[2] Goodfellow et al. Explaining and Harnessing Adversarial Examples. **ICLR**, 2015.

Catastrophic Overfitting

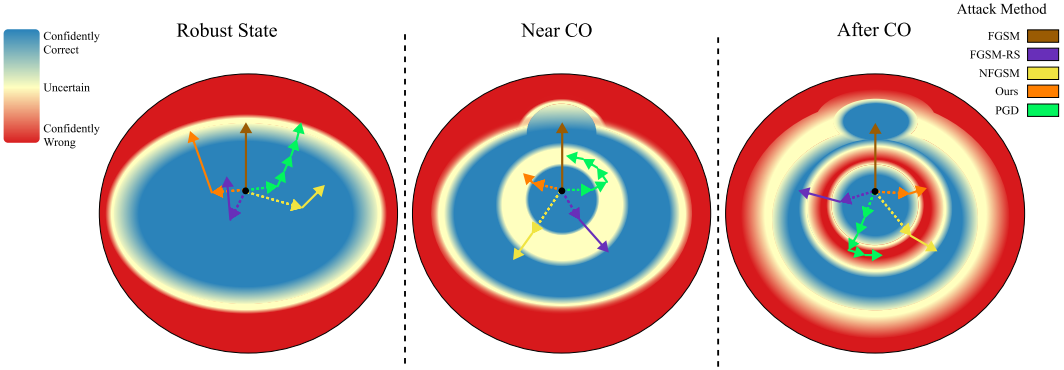
Wong et al.[3] identified a critical failure mode in single-step AT.

- Single-step (FGSM) accuracy rises
- Multi-step (PGD) accuracy drops
- This happens very quickly
- Wong et al.[3] call this **Catastrophic Overfitting**

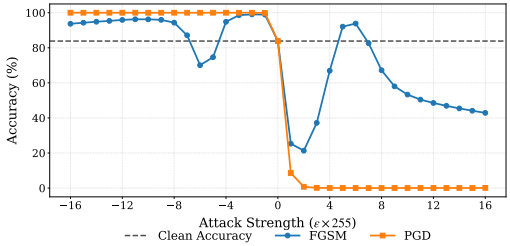


[3] Wong et al. Fast is Better than Free: Revisiting Adversarial Training. **ICLR**, 2020.

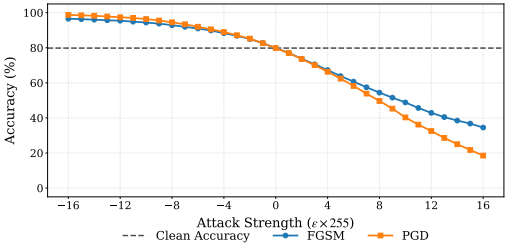
Evolution of the Loss Landscape Geometry



Epsilon Overfitting



(a) FGSM Adversarial Training



(b) SORA Adversarial Training

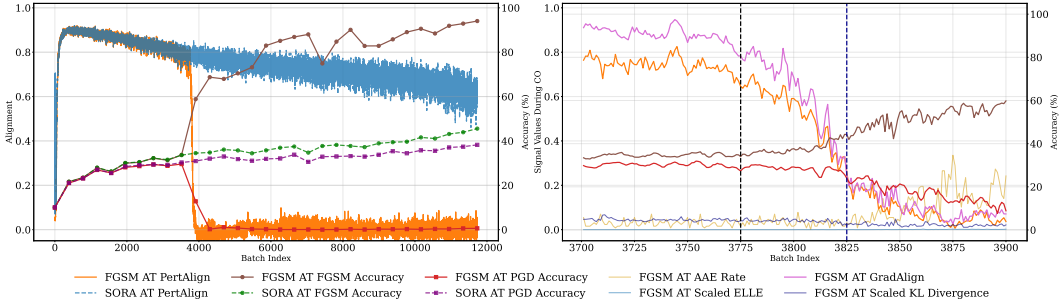
Perturbation Alignment

- Formally, **PertAlign** is defined as:

$$\cos\left(\nabla_{\boldsymbol{x}}\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y), \nabla_{\boldsymbol{x}}\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{\delta}), y)\right)$$

- This measures the alignment between the loss gradient with respect to:
  1. The original sample
  2. The adversarial example

PertAlign in Action



Our General Idea

- If we have **high** **PertAlign**:
  - Loss surface is locally linear
  - Multi- and single-step methods yield similar $\boldsymbol{\delta}$
  - We can increase the step-size
- If we have **low** **PertAlign**:
  - Loss surface is distorted
  - Multi- and single-step methods yield very different $\boldsymbol{\delta}$
  - We must reduce the step-size

Second-Order Adaptive Method (SORA)

- SORA has two main components:
  1. Adaptive maximum step-size:

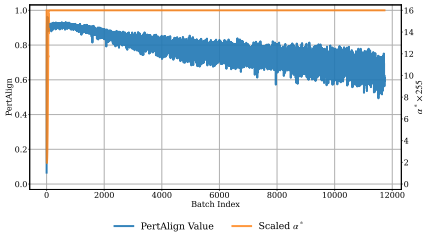
$$\alpha^* = \begin{cases} \min\left(\alpha_{\max}, \frac{\alpha_0}{1 - \frac{p^T g'}{\|g\|_1}}\right), & \frac{p^T g'}{\|g\|_1} < 1, \\ \alpha_{\max}, & \text{otherwise.} \end{cases}$$

  2. Per-pixel channel diversification:

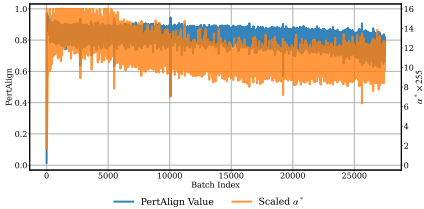
$$\boldsymbol{\alpha} \sim \mathcal{U}\left(0, \alpha^*\right)^d$$

- We can then generate adversarial examples:

$$\boldsymbol{x}' \leftarrow \boldsymbol{x} + \boldsymbol{\eta} + \boldsymbol{\alpha} \odot \operatorname{sign}(\boldsymbol{g})$$

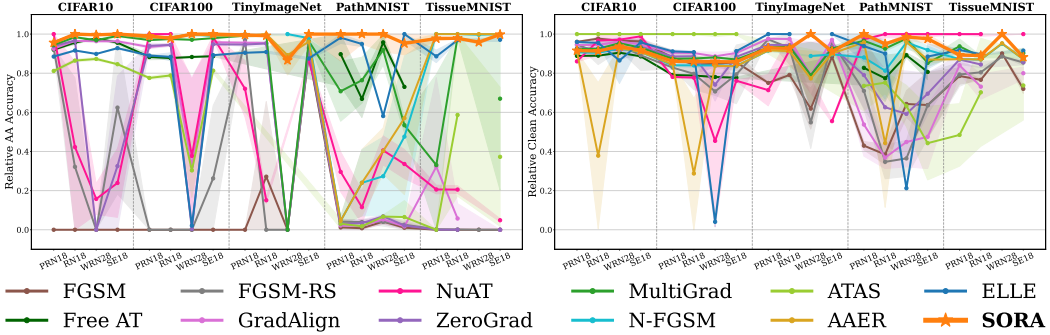


(a) CIFAR-10

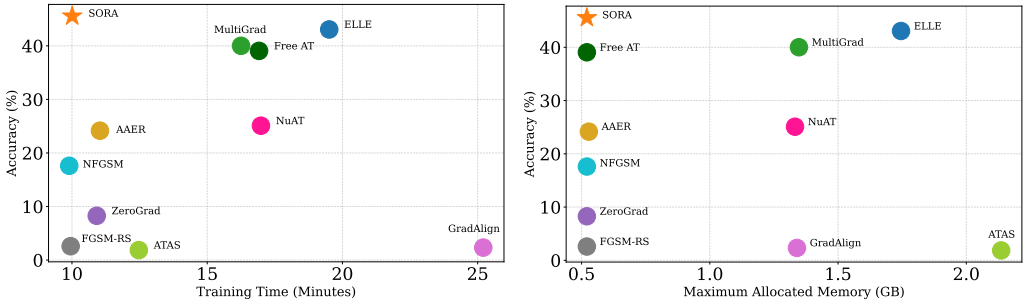


(b) ImageNet-100

Performance and Generalization



Training Costs



(a) Time (b) Memory

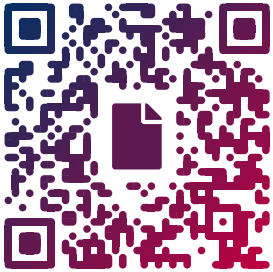
Figure: Training time and memory usage on PathMNIST with PreActResNet-18, trained for 30 epochs, measured on an `NVIDIA GeForce RTX 4090` GPU. The ★ marks SORA. The vertical axis in both figures represents PGD-10 accuracy.

Contributions

- Formalizing the characteristics of a **good solution**:
  - Robustness across datasets and architectures
  - High clean and robust accuracy with low cost
  - Dataset and architecture agnostic hyperparameters
- Understanding CO through the new lens of **Epsilon Overfitting**.
- Introducing **PertAlign** as a CO tracking metric with no overhead.
- Introducing **SORA** as a reliable single-step AT method.

Read Our Paper

- Theoretical insights
- Reinterpreting previous work
- Ablations
- Euclidean norm attacks
- Vision Transformers



Thank you for listening!

Presenters: {mazdak.teymourian01, ramtin.moslemi}@sharif.edu
Corresponding Author: rohban@sharif.edu