Towards Stable and Efficient Adversarial Training against l_1 Bounded Adversarial Attacks

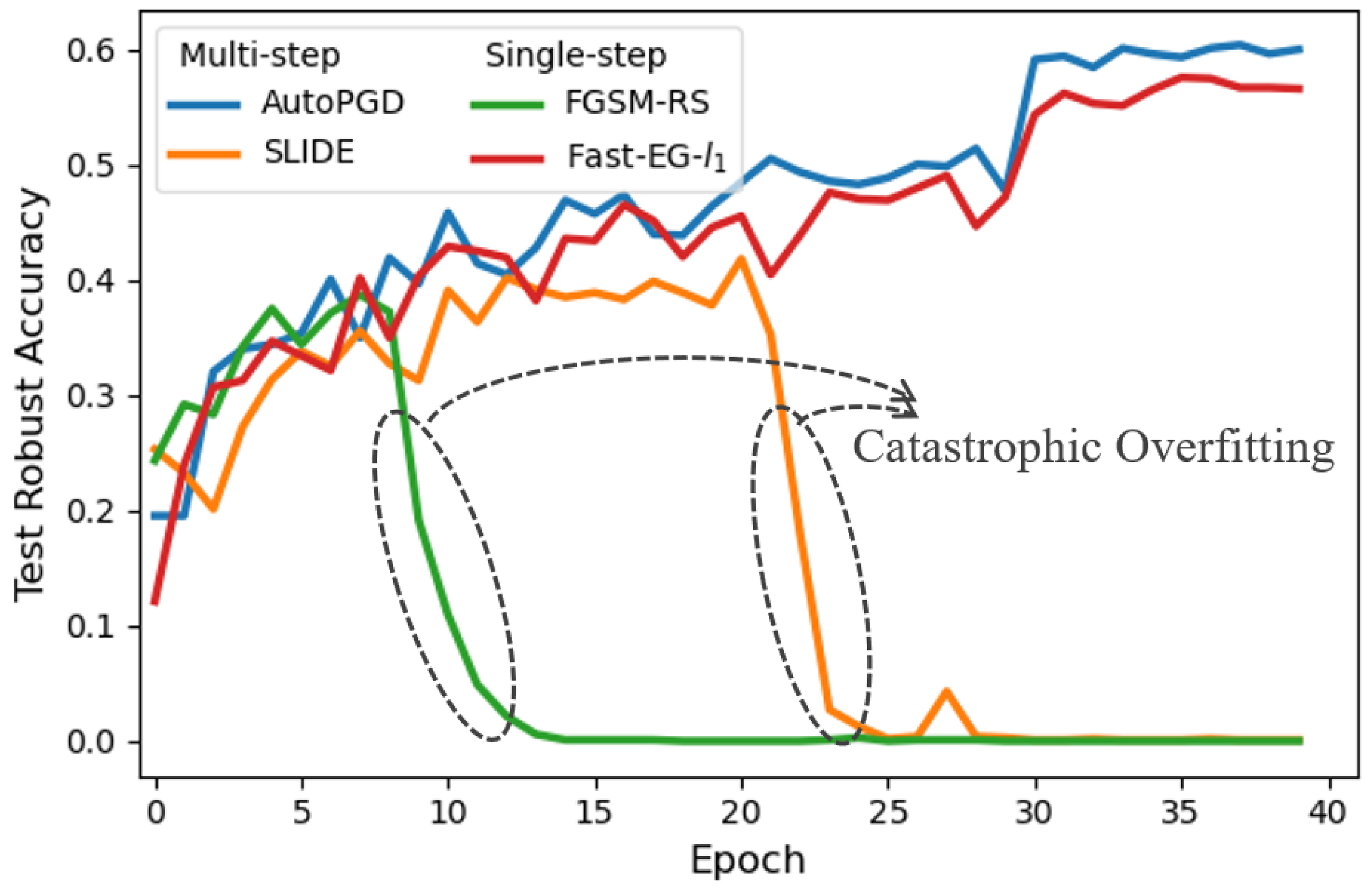
Yulun Jiang*, Chen Liu*, Zhichao Huang, Mathieu Salzmann, Sabine Süsstrunk

EPFL · 香港城市大學 City University of Hong Kong · ByteDance · ICML International Conference On Machine Learning

Contribution

For adversarial training against l_1-norm bounded attacks

- We demonstrate the problem of **catastrophic overfitting** (CO) as a result of overfitting to sparse perturbations.
- We propose **Fast-EG-l_1**, an efficient and stable single-step adversarial training method without CO.



Background

Optimization problem of l_1 adversarial training

$$\min_{\theta} \sum_{i=1}^{N} \max_{\Delta \in \mathcal{S}_{\epsilon}^{(p)}} \mathcal{L}(\theta, \boldsymbol{x}_i + \Delta) \, . \tag{1}$$

with adversarial budget $\mathcal{S}_{\epsilon}^{(p)} := \{\Delta \mid \|\Delta\|_p \leq \epsilon\}$ and $p = 1$.

- Existing methods are based on K**-hot coordinate descent**

$$\Delta \leftarrow \Pi_{\mathcal{S}_{\epsilon}^{(1)}} [\Delta + \alpha / K \cdot \mathbb{1}\{i \in \text{topk}(\nabla \mathcal{L})\}] \tag{2}$$

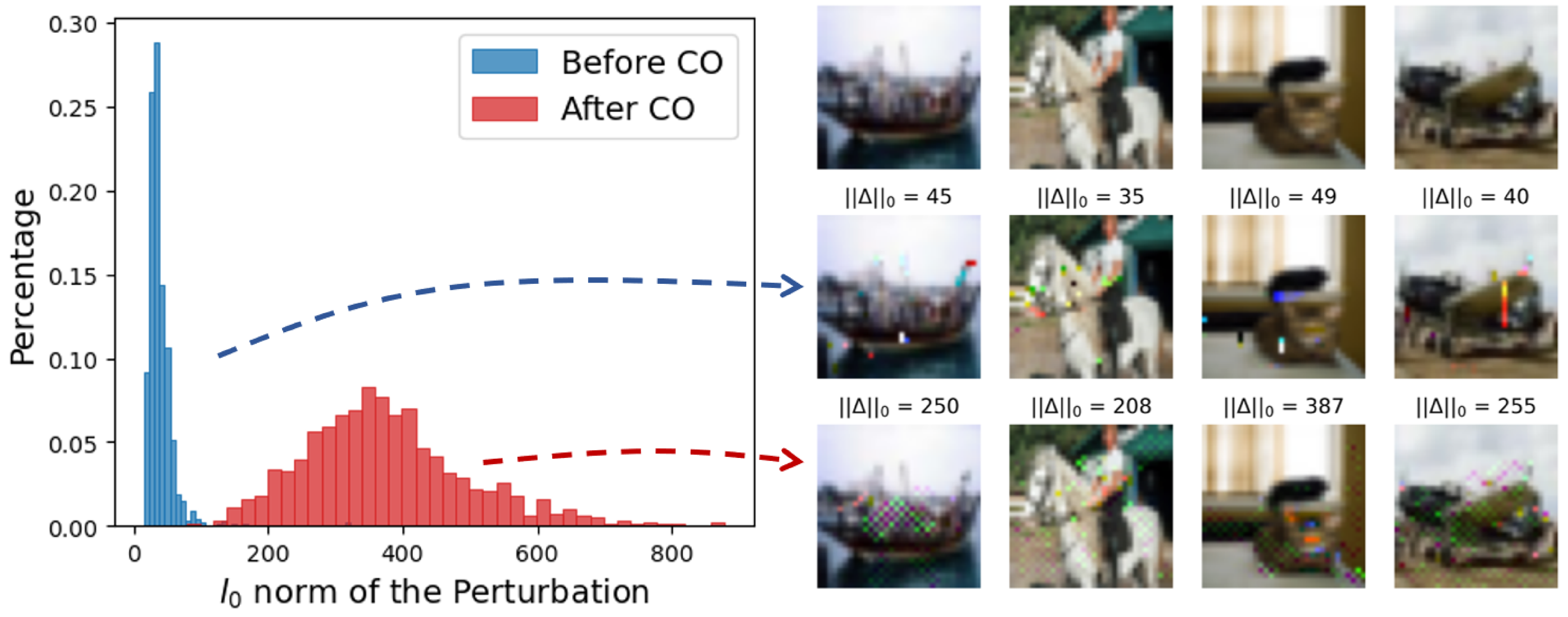
with the problem of efficiency (multi-iters) and stability (CO).

Analysis of CO

Our analysis shows:

- Coordinate descent incurs a strong biased in generating sparse perturbations.
- Model might **overfit to sparse perturbations** and become vulnerable to relatively dense attacks.

→ CO, training **unstable** and **inefficient**.



Method

Our method **Fast-EG-l_1** generates l_1 bounded perturbations based on Euclidean geometry:

$$\Delta \leftarrow \Pi_{\mathcal{S}_{\epsilon_{train}}^{(1)}} (\Delta + \alpha \cdot \nabla \mathcal{L} / \|\nabla \mathcal{L}\|_2) \tag{3}$$

Still project Δ into the l_1-norm budget.

- Setting: larger training budget $\epsilon_{train} \geq \epsilon$, stepsize $\alpha = \sqrt{\epsilon}$.
- Advantages: efficient and stable, w/o CO, no memory overhead or extra hyper-parameters.

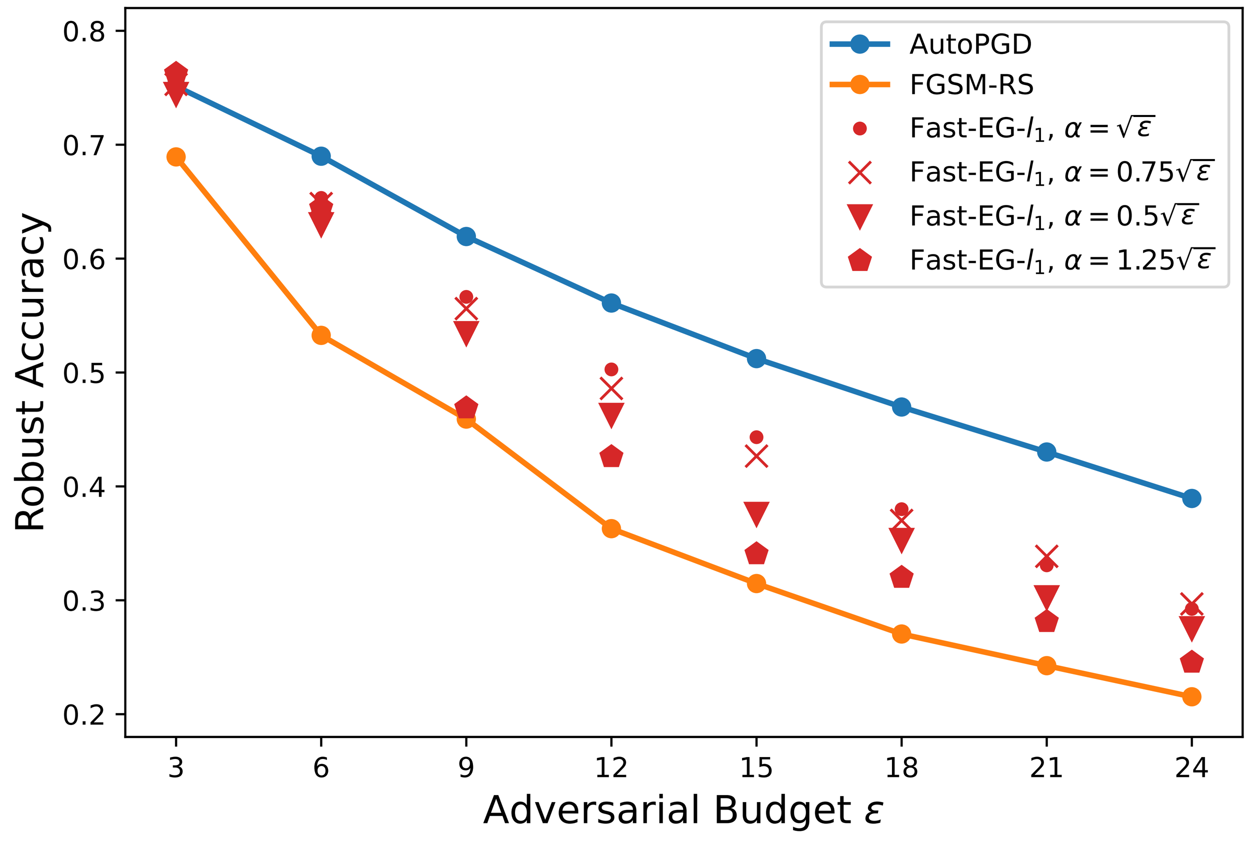
Experiments

Comparison with existing methods

Setting of **Fast-EG-l_1**: $\alpha = \sqrt{\epsilon}$ and $\epsilon_{train} = 2\epsilon$ on all datasets.

Method	CIFAR10 ($\epsilon = 12$) AA (%)	CIFAR10 ($\epsilon = 12$) Time (h)	CIFAR100 ($\epsilon = 6$) AA (%)	CIFAR100 ($\epsilon = 6$) Time (h)	ImageNet100 ($\epsilon = 72$) AA (%)	ImageNet100 ($\epsilon = 72$) Time (h)
AutoPGD	55.77	2.58	42.18	2.58	-	-
FGSM-RS	36.29	0.76	33.23	0.71	36.64	22.12
ATTA	46.57	0.67	33.74	0.68	-	-
AdaAT	31.84	0.88	28.64	0.84	28.62	26.96
Grad-Align	36.38	1.52	33.19	1.52	-	-
N-FGSM	44.21	0.65	35.79	0.66	30.28	23.53
NuAT	48.35	1.01	36.46	1.05	45.82	29.18
Fast-EG-l_1	**50.27**	0.67	**38.03**	0.67	**46.74**	22.11

Ablation Study on α and ϵ (CIFAR10)



Ablation Study on ϵ_{train} with $\epsilon = 12$ (CIFAR10)

ϵ_{train}	ϵ	1.5ϵ	2ϵ	2.5ϵ	3ϵ
Clean (%)	69.70	78.35	76.14	72.77	70.05
Robust (%)	38.15	48.85	**50.27**	49.63	48.10

https://github.com/IVRL/FastAdvL1