Conflict-aware Adversarial Training

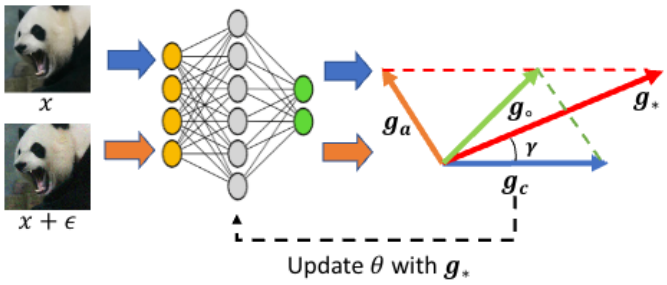
Zhiyu Xue[1], Haohan Wang[2], Yao Qin[1], Ramtin Pedarsani[1]

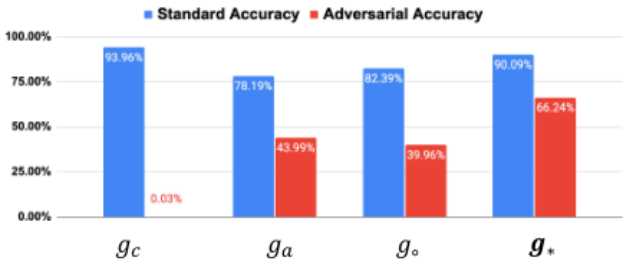
[1]University of California, Santa Barbara

[2]University of Illinois Urbana-Champaign

Introduction

Adversarial training enhances robustness but often sacrifices standard accuracy due to **gradient conflicts** between standard and adversarial losses. **We propose Conflict-Aware Adversarial Training (CA-AT)**, which dynamically balances these losses via gradient operation. CA-AT achieves a better trade-off between accuracy and robustness in both full and parameter-efficient fine-tuning (PEFT) settings.



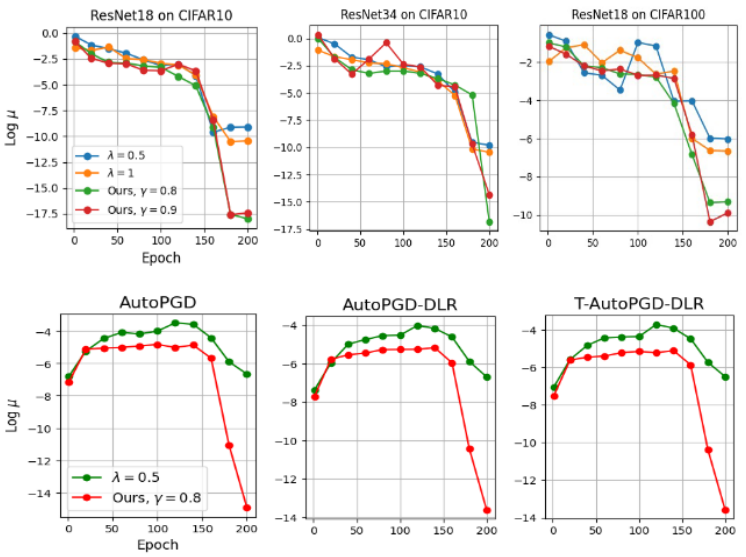


Left: CA-AT resolves the conflict between clean gradient $g_c$and adversarial gradient $g_a$via gradient projection with trade-off factor ϕ. **Right:** The model optimized by g_*achieves higher standard (blue) and adversarial (red) accuracy than the vanilla averaged gradient $g_o$on ResNet18 trained with PGD.

Observation of Gradient Conflict

We define $\mu = ||g_a||_2 \cdot ||g_c|| \cdot (1 - \cos(g_c, g_c))$ to measure gradient conflict.

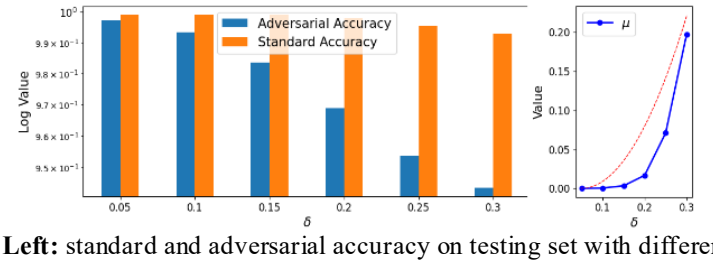
1. $||g_c||_2$ - convergence of standard loss $\mathcal{L}_c$
2. $||g_a||_2$ - convergence of adversarial loss $\mathcal{L}_a$
3. $1 - \cos(g_c, g_c)$ - directional conflict



Theoretical Analysis of Gradient Conflict

Theorem 1. *Consider the gradient conflict* μ *and suppose that the input* x *is a* d*-dimensional vector.*

1. *Given the* L_2*restriction for* ε *as* $||\varepsilon||_2 \leq \delta$ *,we have* $\mu \leq \mathcal{O}(\delta^2)$.
2. *Given the* L_∞*restriction for* ε *as* $||\varepsilon||_\infty \leq \delta$ *,we have* $\mu \leq \mathcal{O}(d^2\delta^2)$.



Left: standard and adversarial accuracy on testing set with different budget δ. **Right:** the relation between ε and δ.

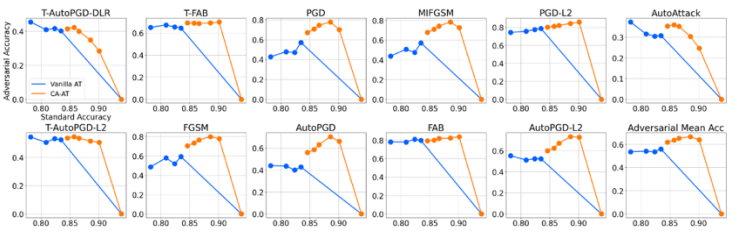
Methodology

$$g_* = \begin{cases} g_a + \dfrac{||g_a||_2(\gamma\sqrt{1-\phi^2} - \phi\sqrt{1-\gamma^2})}{||g_c||_2\sqrt{1-\gamma^2}} g_c, & \phi \leq \gamma \\ g_c, & \phi > \gamma \end{cases}$$

CA-AT introduces a threshold γ to decide whether to use the clean or adversarial gradient. When the similarity is below γ, the method projects g_a toward g_c to mitigate conflict. Otherwise, it directly uses g_c to preserve standard accuracy.

Main Experimental Results

Adversarial Training From Scratch

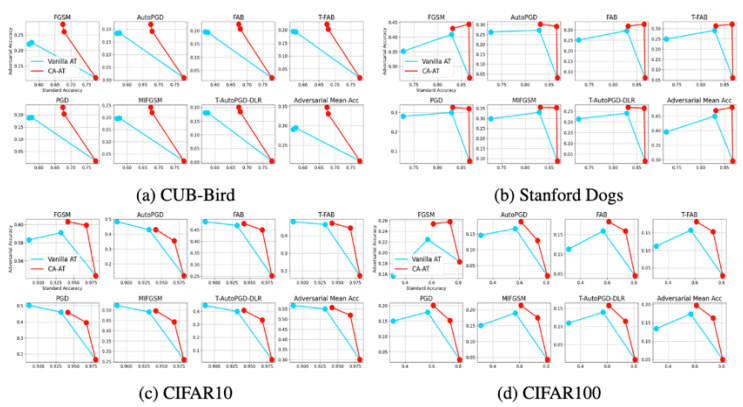


CIFAR10

	Standard	FGSM	PGD	AutoPGD	MIFGSM	T-FAB
Vanilla AT, $\lambda = 0.5$	0.4881	0.1872	0.152	0.1615	0.16	0.347
CA-AT, $\gamma = 0.8$	**0.4989**	**0.254**	**0.1867**	**0.1753**	**0.2044**	**0.3584**

TinyImageNet

Adversarial PEFT



Future Directions

(1) Reveal the influence of training samples causing gradient conflict.
(2) Does gradient conflict exist in adversarial training for LLMs?