CLIP Is Strong Enough to Fight Back: Test-time Counterattacks towards Zero-shot Adversarial Robustness of CLIP

Songlong Xing[1] Zhengyu Zhao[2] Nicu Sebe[1]

[1]University of Trento, Italy [2]Xi'an Jiaotong University, China

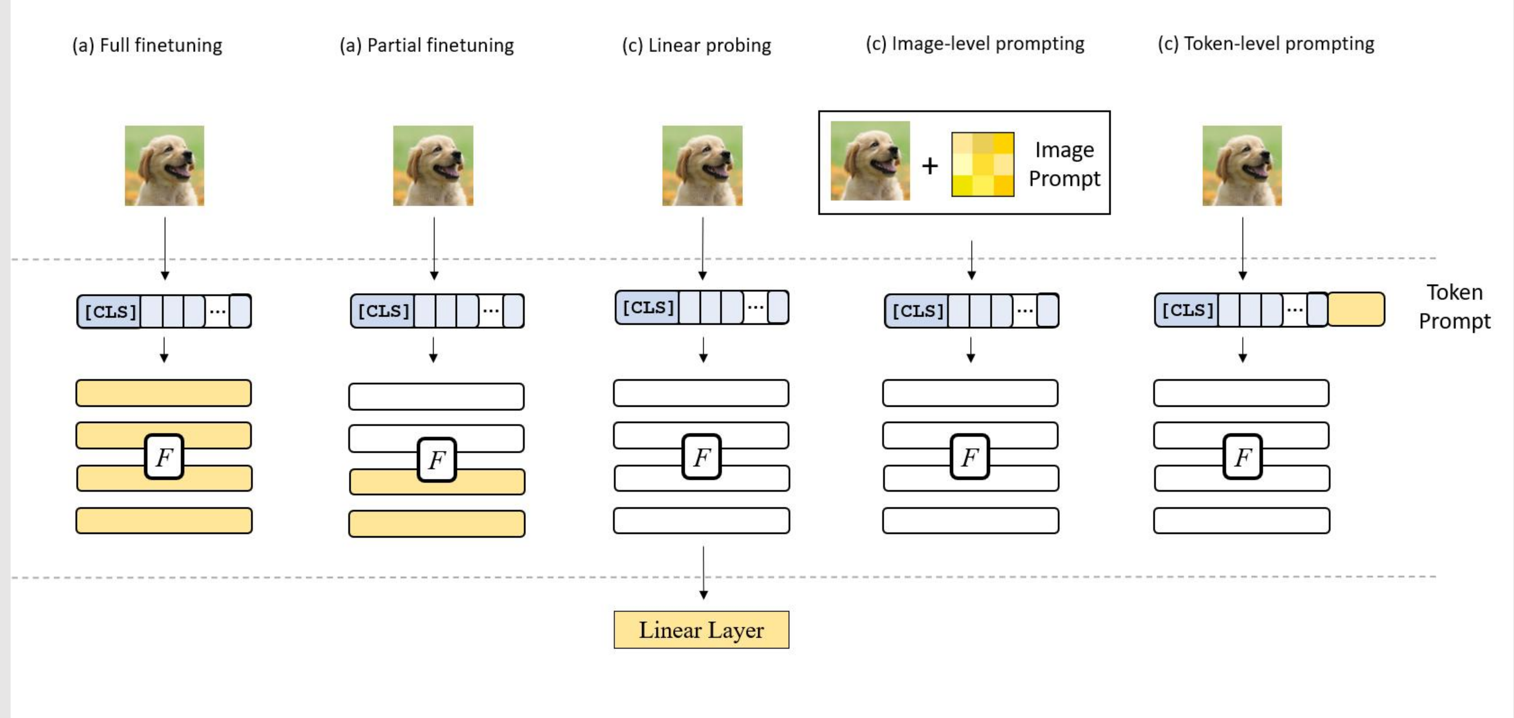
CVPR Nashville JUNE 11-15, 2025

paper & code!

Motivation

1. CLIP is vulnerable to adversarial perturbations

2. State-of-the-art methods employ finetuning or prompt tuning with adversarial images.



(Figure borrowed from TeCoA [1])

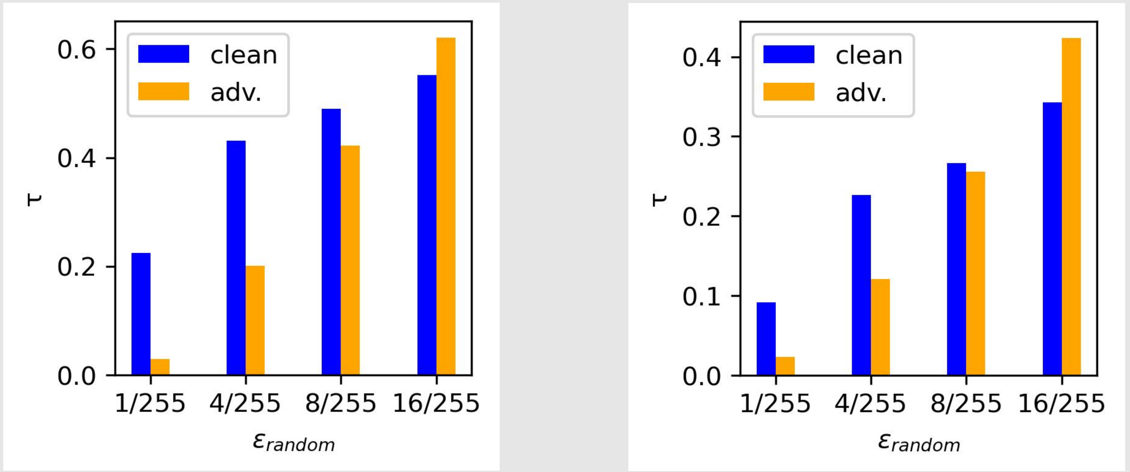
Limitations are apparent:

- Expensive training
- Overfitting to adversarial samples
- Significant loss of clean performance

What if we discard training and counter adversary at test time?

Preliminary Experiment

Perturbations that maximize downstream loss cause 'false robustness'.

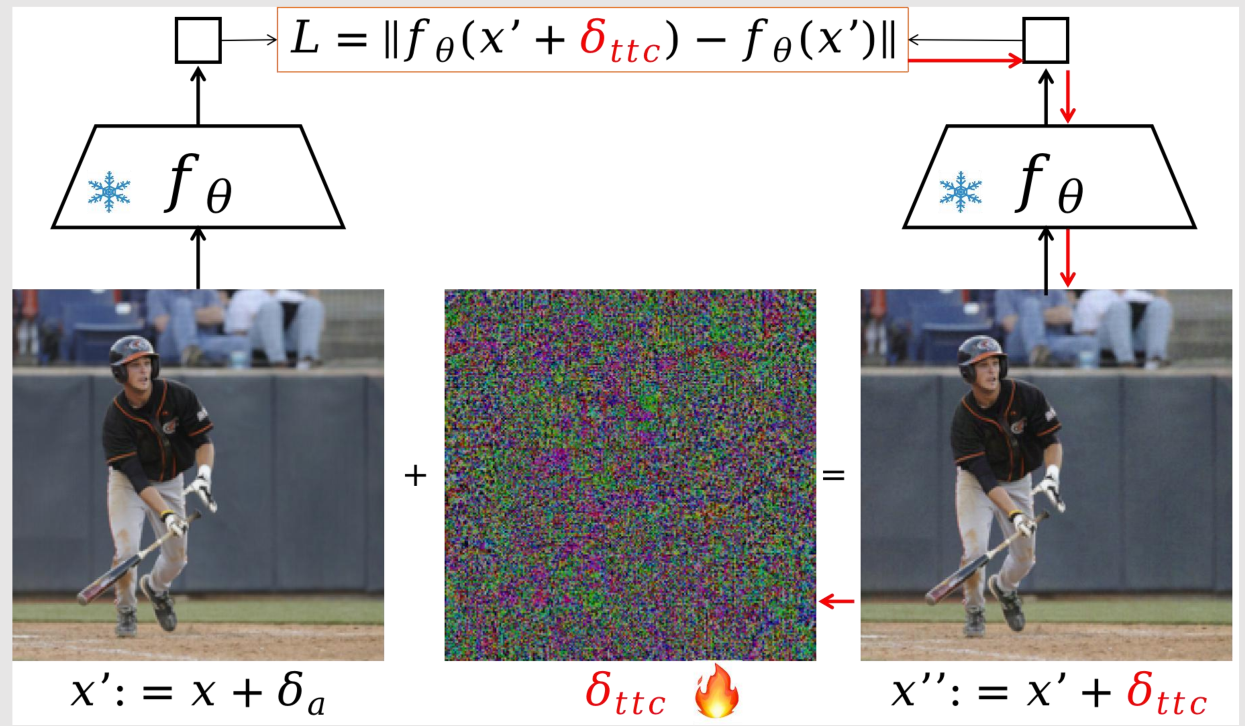


CIFAR100 ImageNet

(Check out Appendix for theoretical explanation)

Methodology

Leverage the vision encoder to counterattack adversarial images to mitigate 'false robustness'.



$$\delta_{ttc} = \arg\max_{\delta} \|f_\theta(x+\delta) - f_\theta(x)\|$$
$$s.t.\ \|\delta\|_\infty \leq \varepsilon_{ttc}$$

Algorithm

1. τ describes representational variation induced by random noise:

$$\tau = \frac{\|f_\theta(x+n) - f_\theta(x)\|}{f_\theta(x)}$$

2. Counterattack test image x depending on τ at the initial step ($\tau < \tau_{thres}$):

$$\delta^i_{ttc} = \prod(\delta^{i-1}_{ttc} + \alpha \nabla_\delta \|f_\theta(x + \delta^{i-1}_{ttc}) - f_\theta(x)\|)$$

3. Weight and sum δ^i_{ttc} across N steps:

$$\delta_{ttc} = \sum_{i=0}^{N} w_i \bullet \delta^i_{ttc}$$

Algorithm 1 τ-thresholded weighted counterattacks.

Require: Test image x, pre-trained CLIP vision encoder f_θ, counterattack budget ϵ_{ttc}, stepsize α, number of steps N, user-defined parameters τ_{thres} and β.

```
procedure TEST-TIME COUNTERATTACKS
    δ⁰_ttc ~ U(−ε_ttc, ε_ttc).
    Compute τ based on Eq. (4) using δ⁰_ttc.
    if τ ≥ τ_thres then
        w_0 = 1
        return δ_ttc = δ⁰_ttc
    else if τ < τ_thres then
        w, δ_ttc := {}, {}
        for i = 1, 2, ..., N do
            δ^i_ttc = Π(δ^{i−1}_ttc + α∇_δ‖f_θ(x+δ^{i−1}_ttc) − f_θ(x)‖)
            w_i = exp(β·i) / Σ_{j=0}^{N} exp(β·j)   (Eq. (5))
            w ← w_i, δ_ttc ← δ^i_ttc
        end for
        return δ_ttc = Σ_{i=0}^{N} w_i · δ^i_ttc   (Eq. (6))
    end if
end procedure
```

Results

(%)		CLIP	TeCoA [1]	FARE [2]	RN	TTC (ours)	Δ (w.r.t. CLIP)
ImageNet	Rob.	1.15	18.89	14.00	1.77	38.41	+37.26
	Acc.	59.69	34.89	48.79	59.34	49.39	-10.30
CIFAR10	Rob.	0.74	33.61	19.65	2.01	28.75	+28.01
	Acc.	85.12	64.61	74.44	81.18	81.18	-3.94
Caltech256	Rob.	8.47	43.19	38.79	11.33	60.11	+51.64
	Acc.	81.72	61.14	73.32	81.25	79.66	-2.06
Cars	Rob.	0.02	8.76	6.75	0.16	33.01	+32.99
	Acc.	52.02	20.91	38.68	52.14	48.16	-3.86
Avg. 16 datasets	Rob.	2.70	26.54	20.00	3.86	39.17	+36.47
	Acc.	61.51	40.25	51.02	61.61	59.75	-1.76

Tab.1 PGD-10 attacks ($\varepsilon_a = 1/255$). Find full table in the paper.

(%)	C10	IN	Cal256	Cars	Rob.	Acc.
TeCoA	33.61	18.89	43.19	8.76	26.54	40.25
TeCoA + TTC	34.68	23.14	48.49	12.09	29.02	39.85
Δ	1.07 ↑	4.25 ↑	5.30 ↑	3.33 ↑	2.48 ↑	−0.40 ↓
FARE	19.65	14.00	38.79	6.85	20.00	51.02
FARE + TTC	35.55	30.52	59.20	20.46	33.89	49.91
Δ	15.90 ↑	16.52 ↑	20.41 ↑	13.61 ↑	13.89 ↑	−1.11 ↓

Tab.2 Applying TTC on finetuned models further improves robustness.

Conclusion & Discussion

1. CLIP can leverage f_θ to counterattack
2. First method to defend CLIP at test time

Limitataions to be addressed:

- limited robustness gains on finetuned models
- Incurs compute expense at test time
- May be circumvented by adaptive attacks

Reference

[1] Mao et al. Understanding zero-shot adversarial robustness for large-scale models. In ICLR, 2023.

[2] Schlarmann et al. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In ICML, 2024.