LORE: Lagrangian-Optimized Robust Embeddings for Visual Encoders

Borna Khodabandeh*[1], Amirabbas Afzali*[2], Amirhossein Afsharrad[1,2],
Seyed Shahabeddin Mousavi[1,2], Sanjay Lall[1], Sajjad Amini[3], Seyed-Mohsen Moosavi-Dezfooli[4]

[1]Stanford University, [2]Aktus AI, [3]University of Massachusetts Amherst, [4]Apple

Motivation

Challenges in Visual Encoders:

- Despite their success, visual encoders are still vulnerable to adversarial perturbations.

Motivation

Challenges in Visual Encoders:

- Despite their success, visual encoders are still vulnerable to adversarial perturbations.
- Existing unsupervised adversarial fine-tuning methods show **unstable training** and an unfavorable **robustness–accuracy trade-off**.

Preliminary

FARE [1] proposes unsupervised fine-tuning of the CLIP vision encoder by aligning clean and adversarial embeddings:

$$\mathcal{L}_{\text{FARE}}(\phi_\theta, x) = \max_{\delta:\|\delta\|_\infty \leq \varepsilon} \left\| \phi_\theta(x+\delta) - \phi_{\theta_0}(x) \right\|_2^2.$$

Preliminary

FARE [1] proposes unsupervised fine-tuning of the CLIP vision encoder by aligning clean and adversarial embeddings:

$$\mathcal{L}_{\text{FARE}}(\phi_\theta, x) = \max_{\delta: \|\delta\|_\infty \leq \varepsilon} \left\| \phi_\theta(x+\delta) - \phi_{\theta_0}(x) \right\|_2^2.$$

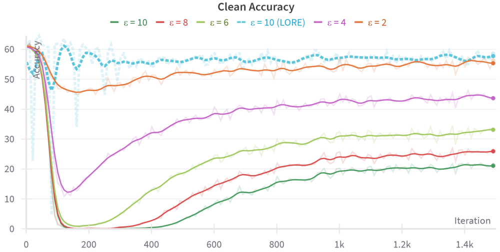
What is the problem?

FARE [1] proposes unsupervised fine-tuning of the CLIP vision encoder by aligning clean and adversarial embeddings:

$$\mathcal{L}_{\text{FARE}}(\phi_\theta, x) = \max_{\delta:\|\delta\|_\infty \leq \varepsilon} \left\|\phi_\theta(x+\delta) - \phi_{\theta_0}(x)\right\|_2^2.$$

What is the problem?

(i) Early accuracy degradation



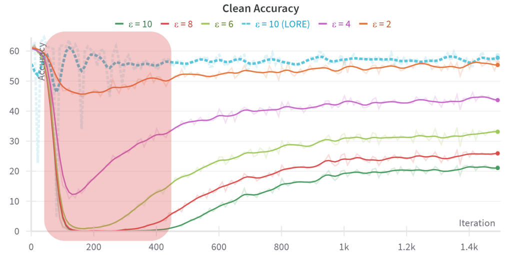
Clean accuracy degradation under different perturbations.

FARE [1] proposes unsupervised fine-tuning of the CLIP vision encoder by aligning clean and adversarial embeddings:

$$\mathcal{L}_{\text{FARE}}(\phi_\theta, x) = \max_{\delta:\|\delta\|_\infty \leq \varepsilon} \left\| \phi_\theta(x+\delta) - \phi_{\theta_0}(x) \right\|_2^2.$$

What is the problem?

(i) Early accuracy degradation



Clean accuracy degradation under different perturbations.

Preliminary

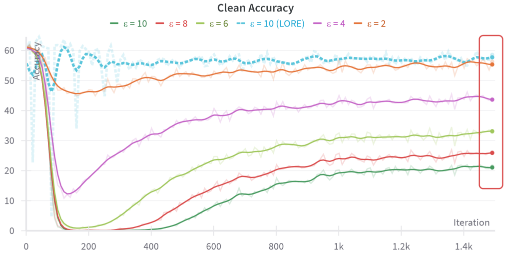
FARE [1] proposes unsupervised fine-tuning of the CLIP vision encoder by aligning clean and adversarial embeddings:

$$\mathcal{L}_{\text{FARE}}(\phi_\theta, x) = \max_{\delta:\|\delta\|_\infty \leq \varepsilon} \left\|\phi_\theta(x+\delta) - \phi_{\theta_0}(x)\right\|_2^2.$$

What is the problem?

(i) Early accuracy degradation

- Significant drop in clean accuracy at the convergence point.



Clean accuracy degradation under different perturbations.

Preliminary

What is the problem?

(i) Early accuracy degradation

What is the problem?

(i) Early accuracy degradation

A simple solution:

Naively adding a regularization term helps preserve clean accuracy:

$$\mathcal{L}_{\text{FARE}-\text{reg}}(\phi_\theta, x) = \max_{\delta:\|\delta\|_\infty \leq \varepsilon} \left\|\phi_\theta(x+\delta) - \phi_{\theta_0}(x)\right\|_2^2 + \lambda\|\phi_\theta(x) - \phi_{\text{org}}(x)\|_2^2,$$

Preliminary

What is the problem?

(i) Early accuracy degradation

A simple solution:

Naively adding a regularization term helps preserve clean accuracy:

$$\mathcal{L}_{\text{FARE}-\text{reg}}(\phi_\theta, x) = \max_{\delta:\|\delta\|_\infty \leq \varepsilon} \left\|\phi_\theta(x+\delta) - \phi_{\theta_0}(x)\right\|_2^2 + \lambda\|\phi_\theta(x) - \phi_{\text{org}}(x)\|_2^2,$$

This time, what is the problem?

Preliminary

What is the problem?

(i) Early accuracy degradation

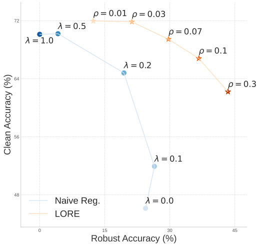
A simple solution:

Naively adding a regularization term helps preserve clean accuracy:

$$\mathcal{L}_{\text{FARE}-\text{reg}}(\phi_\theta, x) = \max_{\delta:\|\delta\|_\infty \leq \varepsilon} \left\|\phi_\theta(x+\delta) - \phi_{\theta_0}(x)\right\|_2^2 + \lambda\|\phi_\theta(x) - \phi_{\text{org}}(x)\|_2^2,$$

This time, what is the problem?

(ii) Practical ineffectiveness of naive regularization



Robustness-accuracy trade-off.

Preliminary

What is the problem?

(i) Early accuracy degradation

A simple solution:

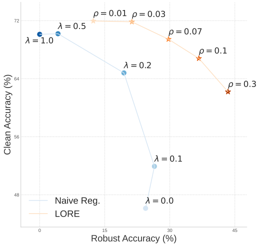
Naively adding a regularization term helps preserve clean accuracy:

$$\mathcal{L}_{\text{FARE}-\text{reg}}(\phi_\theta, x) = \max_{\delta:\|\delta\|_\infty \le \varepsilon} \left\|\phi_\theta(x+\delta) - \phi_{\theta_0}(x)\right\|_2^2 + \lambda\|\phi_\theta(x) - \phi_{\text{org}}(x)\|_2^2,$$

This time, what is the problem?

(ii) Practical ineffectiveness of naive regularization

- It introduces a steep robustness trade-off.



Robustness-accuracy trade-off.

- **Main idea.** Unsupervised extension of constrained optimization, keeping the fine-tuned encoder close to the pre-trained model.

- **Main idea.** Unsupervised extension of constrained optimization, keeping the fine-tuned encoder close to the pre-trained model.

- This yields a **semi-infinite constrained** objective that balances robustness and nominal performance stability, as formulated in:

$$\begin{aligned} &\min_{\theta \in \Theta} \quad \mathbb{E}_{x \sim \mathcal{D}} \left[\max_{\delta \in \Delta} d(\phi_\theta(x+\delta), \phi_{\theta_0}(x)) \right], \\ &\text{s.t.} \quad d(\phi_\theta(x), \phi_{\theta_0}(x)) \leq \rho\, m(x), \quad \text{for almost every } x \in \mathcal{D}. \end{aligned} \tag{1}$$

LORE: Lagrangian-Optimized Robust Embeddings

- **Main idea.** Unsupervised extension of constrained optimization, keeping the fine-tuned encoder close to the pre-trained model.

- This yields a **semi-infinite constrained** objective that balances robustness and nominal performance stability, as formulated in:

$$\begin{aligned} \min_{\theta \in \Theta} \quad & \mathbb{E}_{x \sim \mathcal{D}} \left[\max_{\delta \in \Delta} d(\phi_\theta(x+\delta), \phi_{\theta_0}(x)) \right], \\ \text{s.t.} \quad & d(\phi_\theta(x), \phi_{\theta_0}(x)) \leq \rho\, m(x), \quad \text{for almost every } x \in \mathcal{D}. \end{aligned} \tag{1}$$

- How to handle infinite constraints? $\rightarrow$ Functional Lagrangian

Solving the Constrained Problem

- We employ **Lagrangian duality** to approximate the solution:

$$\max_{\omega\in\Omega}\min_{\theta\in\Theta}\ \mathbb{E}_{x\sim\mathcal{D}}\left[\max_{\delta\in\Delta}\|\phi_\theta(x+\delta)-\phi_{\theta_0}(x)\|_2^2+\lambda_\omega(x)\left(\|\phi_\theta(x)-\phi_{\theta_0}(x)\|_2^2-\rho\|\phi_{\theta_0}(x)\|_2^2\right)\right]. \quad (2)$$

Solving the Constrained Problem

- We employ **Lagrangian duality** to approximate the solution:

$$\max_{\omega \in \Omega} \min_{\theta \in \Theta} \mathbb{E}_{x \sim \mathcal{D}} \left[\max_{\delta \in \Delta} \|\phi_\theta(x+\delta) - \phi_{\theta_0}(x)\|_2^2 + \lambda_\omega(x) \left(\|\phi_\theta(x) - \phi_{\theta_0}(x)\|_2^2 - \rho \|\phi_{\theta_0}(x)\|_2^2 \right) \right]. \quad (2)$$

- **Optimization.** During training, adversarial samples are generated for each batch, followed by K primal updates of encoder parameters θ and one dual update of ω.

Results

1. Controlling the Robustness–Accuracy Trade-off
2. Out-of-Distribution Robustness
3. Image Classification
    - *Zero-shot Image Classification*
    - *In-domain Image Classification*
    - *Robustness at High Adversarial Intensity*

Results

1. Controlling the Robustness–Accuracy Trade-off

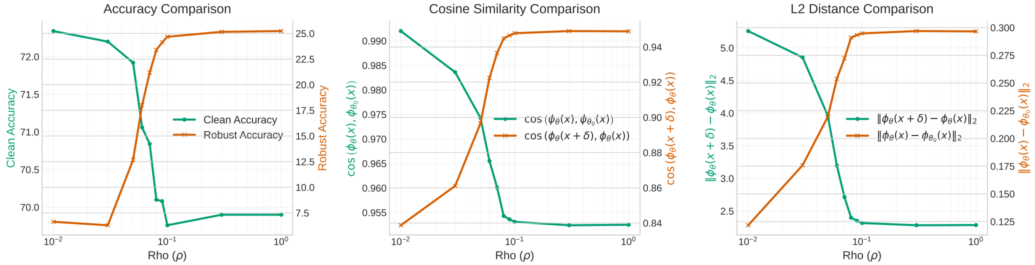


Figure 1: Influence of constraint threshold ρ on model behavior. As ρ increases, robustness improves at the cost of clean data accuracy, cosine alignment, and embedding fidelity, highlighting the effectiveness of controlling the trade-off between robustness and fidelity by tuning ρ in LORE.

2. Out-of-Distribution Robustness

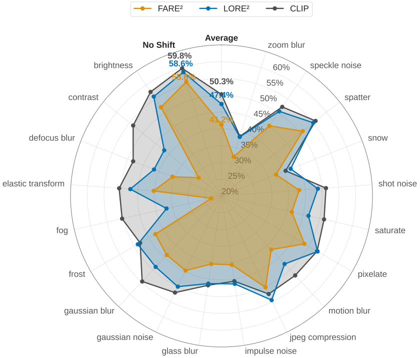


Figure 2: Robustness to common corruptions on ImageNet-C as an OOD evaluation.

Results

3. Zero-shot Image Classification

Table 3: A comprehensive evaluation of clean and adversarial performance is conducted across various image classification datasets using the ViT-B/32 CLIP model. All models are trained on ImageNet and evaluated in a zero-shot setting across diverse benchmarks. Our method consistently achieves a performance increase (↑) relative to the corresponding FARE models.

Eval.	Vision encoder		Zero-shot datasets													Average Zero-shot	
		ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10		
clean	CLIP	59.8	84.1	59.6	89.7	63.3	44.4	46.1	19.6	66.3	69.3	42.3	62.3	87.5	97.2	64.0	
	FARE^1	56.6	84.0	56.3	86.4	61.1	40.5	27.2	18.1	62.0	66.4	40.5	55.5	86.1	95.8	60.0	
	LORE^1	57.4	84.4	55.9	88.5	64.5	40.1	29.9	16.7	61.3	67.2	41.5	53.8	86.9	96.3	60.5	↑0.5
	FARE^2	52.9	82.2	49.7	76.3	51.1	36.4	18.4	15.7	53.3	60.4	35.9	48.2	82.7	93.0	54.1	
	LORE^2	55.7	83.0	51.0	83.4	59.7	37.2	23.0	15.9	54.5	63.4	39.3	51.2	84.3	94.5	57.0	↑2.9
	FARE^4	42.6	78.1	36.5	55.9	35.8	28.8	15.7	10.6	36.1	49.3	27.1	50.0	71.8	85.6	44.7	
	LORE^4	50.1	80.3	40.1	72.4	49.6	32.4	17.7	11.4	39.7	55.1	33.6	50.0	79.3	90.4	50.2	↑5.5
$\varepsilon = 1.0$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	
	FARE^1	27.8	68.6	16.1	61.0	35.6	22.5	6.1	2.9	30.6	34.4	22.5	24.7	55.8	82.2	35.6	
	LORE^1	32.9	71.0	18.7	67.1	40.0	23.7	9.4	4.2	33.5	37.6	24.8	28.3	60.5	84.1	38.7	↑3.1
	FARE^2	34.3	75.2	22.6	60.1	35.4	24.7	12.6	5.3	33.9	39.7	24.1	30.4	64.8	83.3	39.4	
	LORE^2	39.3	76.3	23.3	67.0	43.2	26.4	12.3	6.5	35.8	42.4	26.4	39.0	68.5	85.6	42.5	↑3.1
	FARE^4	33.2	74.8	21.4	44.9	28.0	22.4	14.0	5.8	27.3	37.1	21.3	50.2	59.3	77.7	37.2	
	LORE^4	41.8	77.2	24.1	61.2	39.9	24.5	14.3	7.8	30.2	41.6	25.5	50.2	68.8	83.2	42.2	↑5.0
$\varepsilon = 2.0$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	
	FARE^1	8.0	43.5	1.9	31.0	14.7	12.9	0.6	0.2	6.8	13.4	11.7	14.1	15.9	54.9	17.0	
	LORE^1	13.1	49.0	3.3	37.9	19.0	14.2	2.5	0.5	10.1	17.6	13.1	19.1	23.1	61.2	20.8	↑3.8
	FARE^2	19.3	59.9	7.7	41.2	22.8	17.8	9.6	1.5	16.4	24.2	15.9	23.4	38.6	68.6	26.7	
	LORE^2	24.0	63.3	8.6	47.2	27.2	18.2	10.6	1.7	18.5	26.0	18.4	28.0	44.4	73.1	29.6	↑2.9
	FARE^4	24.1	65.5	10.4	36.0	21.6	18.8	12.3	2.7	17.9	27.7	15.8	50.0	44.4	68.8	30.1	
	LORE^4	32.6	69.5	12.4	50.8	29.6	20.9	13.0	3.3	21.6	32.3	20.0	50.1	55.9	76.1	35.0	↑4.9
$\varepsilon = 4.0$	CLIP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	
	FARE^1	0.3	6.3	0.0	1.7	2.0	2.3	0.0	0.0	0.1	2.6	2.4	0.9	0.0	5.3	1.8	
	LORE^1	0.7	9.7	0.0	3.5	3.1	4.0	0.0	0.0	0.2	3.8	2.8	2.7	0.0	9.4	3.0	↑1.2
	FARE^2	3.2	27.5	0.5	12.3	7.0	7.7	4.3	0.0	2.4	6.8	5.1	15.8	3.0	30.1	9.4	
	LORE^2	5.7	31.1	0.7	13.0	8.2	9.7	0.8	0.0	3.1	8.3	6.5	18.2	7.2	33.5	10.8	↑1.4
	FARE^4	10.7	46.3	1.5	19.7	11.8	11.9	10.2	0.6	6.4	11.4	8.7	45.2	16.2	46.1	18.2	
	LORE^4	17.8	54.2	2.8	27.4	16.8	14.4	10.0	0.6	8.0	16.4	11.7	48.4	25.5	56.1	22.5	↑4.3

3. In-domain Image Classification

Table 1: Clean and adversarial accuracy for in-domain image classification on ImageNet-100 across different CLIP vision encoders, evaluated using the APGD attack.

Method	Backbone	Clean	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
FARE[2]	ViT-B/16	70.40	53.0	34.9	8.8	0.06
LORE[2]	ViT-B/16	**74.7**	62.3	47.7	20.8	0.74
FARE[4]	ViT-B/16	58.1	47.7	37.1	19.0	2.22
LORE[4]	ViT-B/16	71.5	**62.3**	**53.3**	**34.7**	**9.06**
FARE[2]	ViT-B/32 LAION	65.4	41.0	19.0	2.02	0.02
LORE[2]	ViT-B/32 LAION	**70.2**	**51.8**	**31.4**	7.26	0.04
FARE[4]	ViT-B/32 LAION	52.7	36.7	23.4	6.72	0.20
LORE[4]	ViT-B/32 LAION	68.4	44.7	29.6	**10.7**	**0.62**
FARE[2]	ConvNeXt-B	74.2	61.6	46.1	16.7	0.22
LORE[2]	ConvNeXt-B	**75.6**	64.9	52.4	25.6	1.04
FARE[4]	ConvNeXt-B	70.6	61.6	52.3	32.7	6.48
LORE[4]	ConvNeXt-B	73.5	**66.0**	**58.1**	**40.3**	**10.4**

Table 2: Clean and adversarial accuracy for in-domain image classification on ImageNet across different DINOv2 variants. Adversarial robustness is evaluated using APGD attack.

Method	Backbone	Clean	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
FARE[4]	ViT-S/14	69.2	60.7	**51.2**	30.7	2.91
LORE[4]	ViT-S/14	**77.3**	**60.8**	50.0	30.3	5.8
FARE[8]	ViT-S/14	55.1	48.9	42.7	30.0	8.13
LORE[8]	ViT-S/14	75.1	55.9	48.8	**36.8**	**13.7**
FARE[4]	ViT-B/14	78.3	71.9	64.1	44.0	6.51
LORE[4]	ViT-B/14	80.2	**73.5**	**67.1**	**49.6**	11.2
FARE[8]	ViT-B/14	69.4	63.8	57.8	44.1	16.0
LORE[8]	ViT-B/14	**80.5**	65.0	59.7	48.5	**21.8**

3. Robustness at High Adversarial Intensity

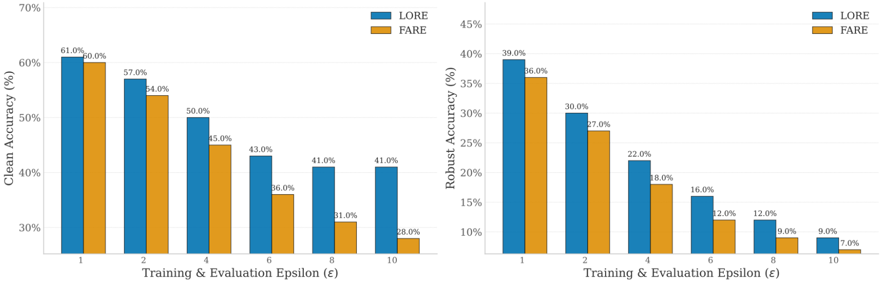
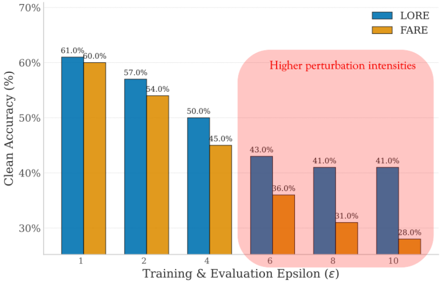


Figure 5: Comparison of LORE and FARE across different training and evaluation perturbations (ε). LORE consistently outperforms FARE, particularly at higher ε values, achieving higher robust accuracy while maintaining better clean performance, especially at higher perturbation intensities.

3. Robustness at High Adversarial Intensity



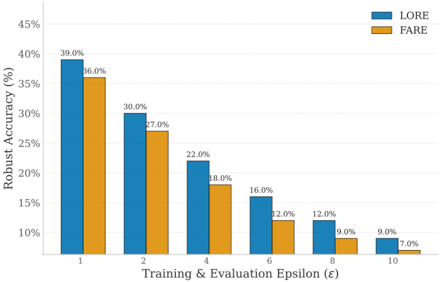


Figure 5: Comparison of LORE and FARE across different training and evaluation perturbations (ε). LORE consistently outperforms FARE, particularly at higher ε values, achieving higher robust accuracy while maintaining better clean performance, **especially at higher perturbation intensities**.

References

[1] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models, 2024. URL https://arxiv.org/abs/2402.12336.