Robustness Verification for Contrastive Learning

Zekai Wang, Weiwei Liu

School of Computer Science, Wuhan University, China

ICML 2022, Accept for Long Presentation

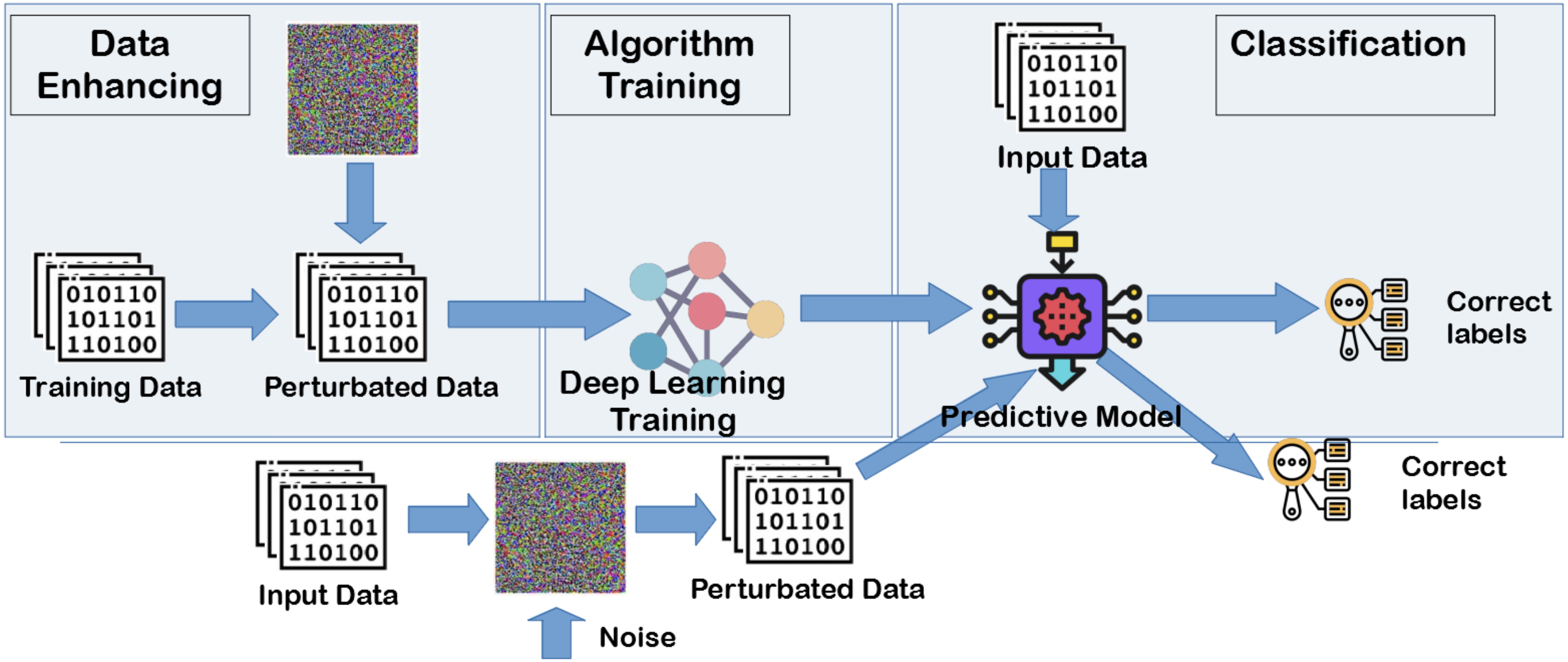
Content

- Background
- Motivation
- RVCL Framework
- Experiments

Background: Adversarial Training

- Define the perturbation: $\delta = \arg\max_{\|\delta'\|_\infty \leq \epsilon} \ell(\theta, x + \delta')$
- Adversarial training aims to solve the optimization problem:

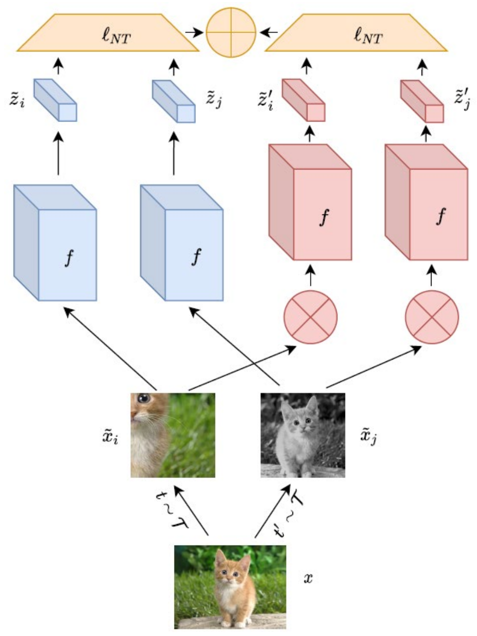
$$\min_{\theta} \; \mathbb{E}_{x \in \mathcal{X}} \, \ell(\theta, x + \delta)$$



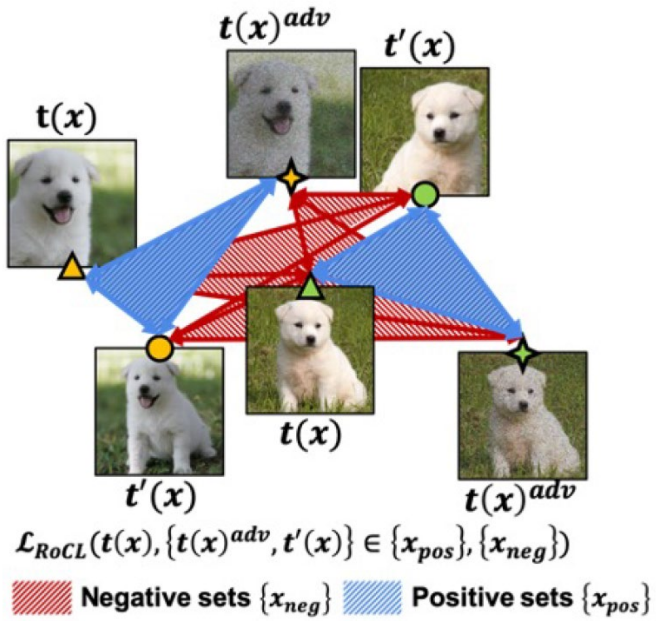
https://towardsdatascience.com/adversarial-machine-learning-mitigation-adversarial-learning-9ae04133c137

Background: Contrastive Adversarial Training

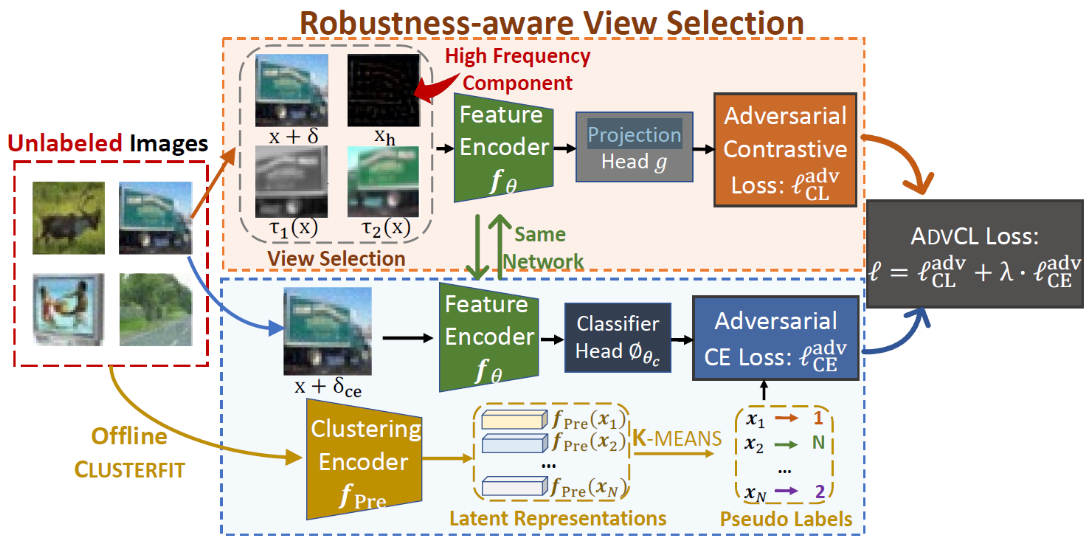
- Labeling scarcity amplified in adversarial robust training
  - ➢ Sample complexity is significantly higher than standard training
- Prior works explore using unlabeled data to generate robust models
  - ➢ Combine adversarial training with **contrastive learning**



ACL, Jiang et al., 2020



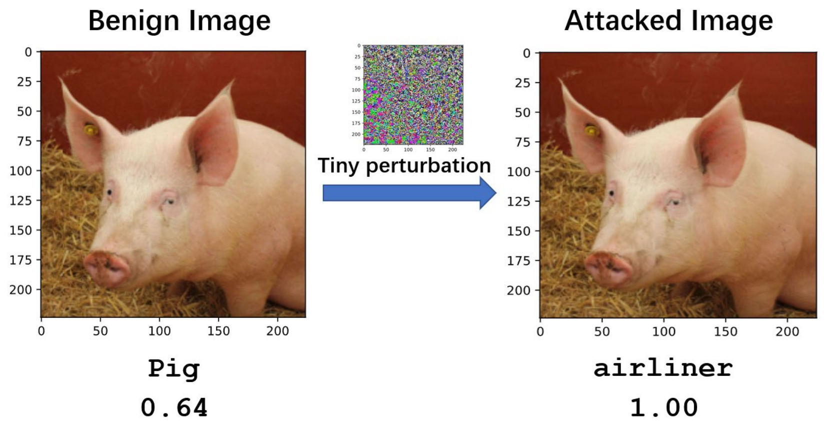
ROCL, Kim et al., 2020



AdvCL, Fan et al., 2021

Motivation

- Existing contrastive AT methods use the empirical robustness metric to evaluate the robustness of encoders, an approach that relies on **attack algorithms**, **image labels** and **downstream tasks**



attack algorithms

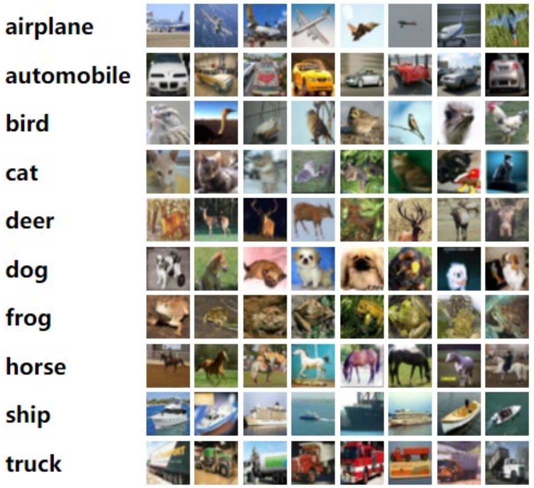
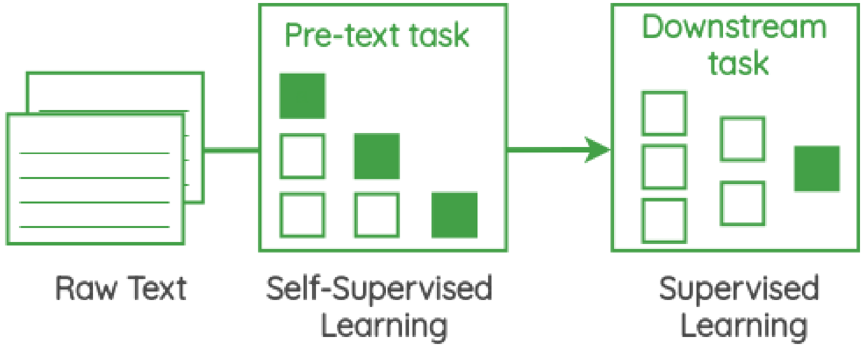


image labels

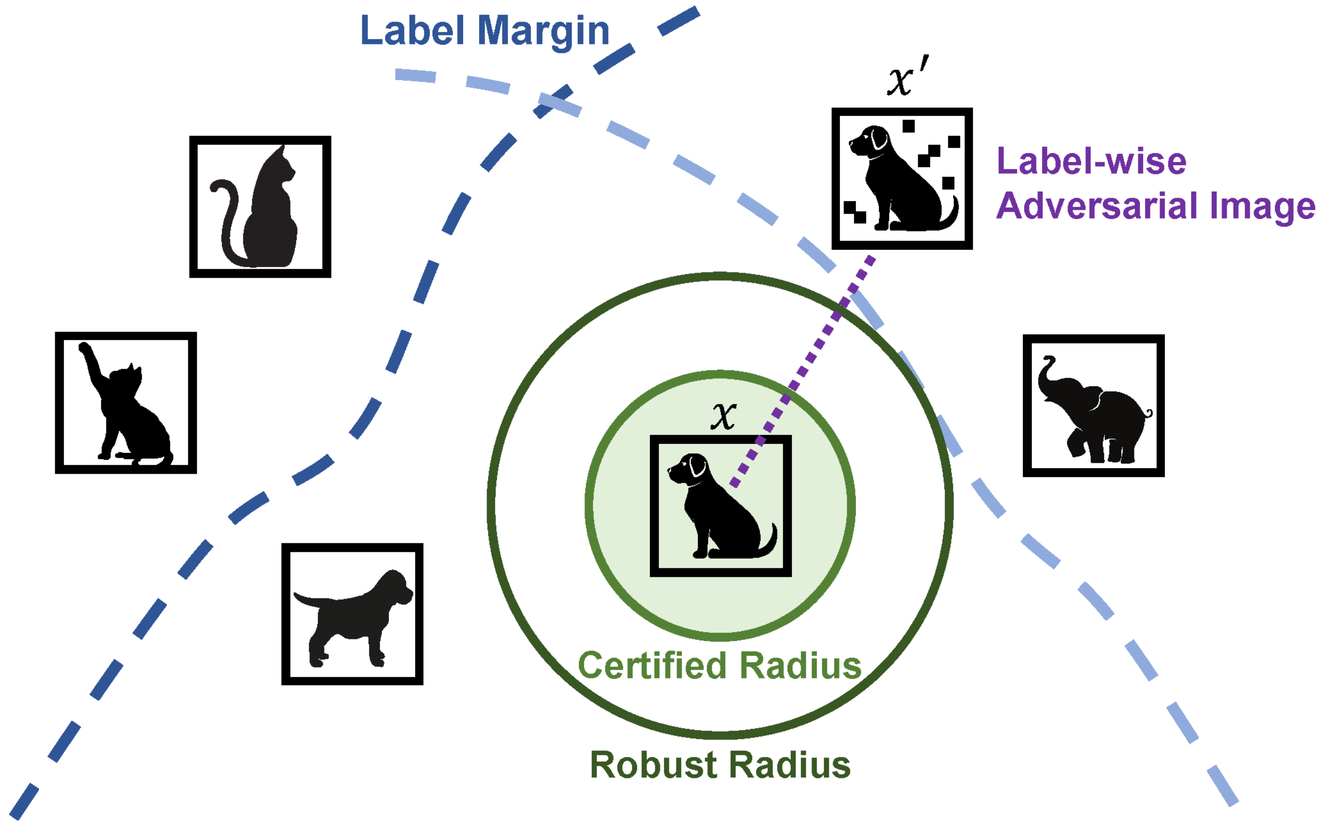


downstream tasks

https://amitness.com/2020/05/self-supervised-learning-nlp

Background: Supervised Robustness Verification

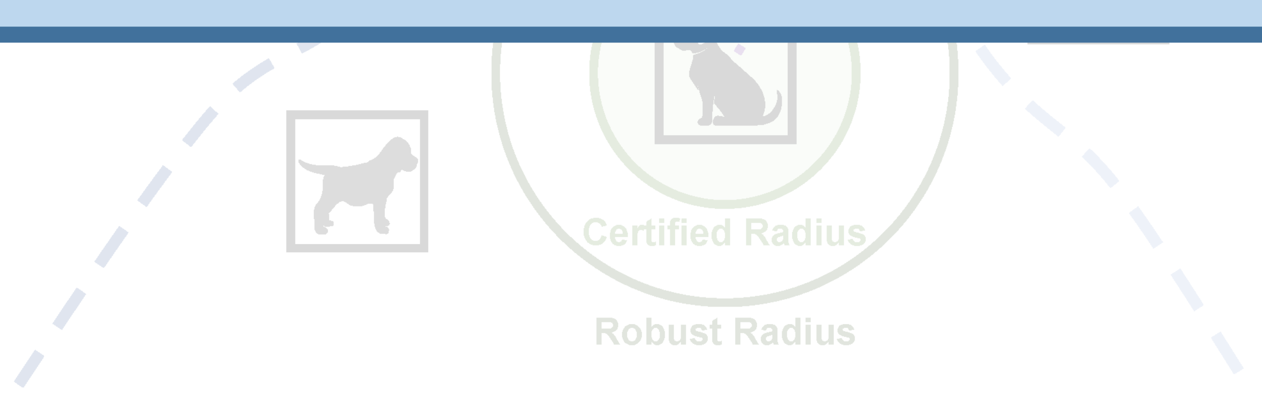
- **Robustness verification** means classifiers whose prediction at point x is verified to be constant within a neighborhood of x, regardless of what attack algorithm is applied



Background: Supervised Robustness Verification

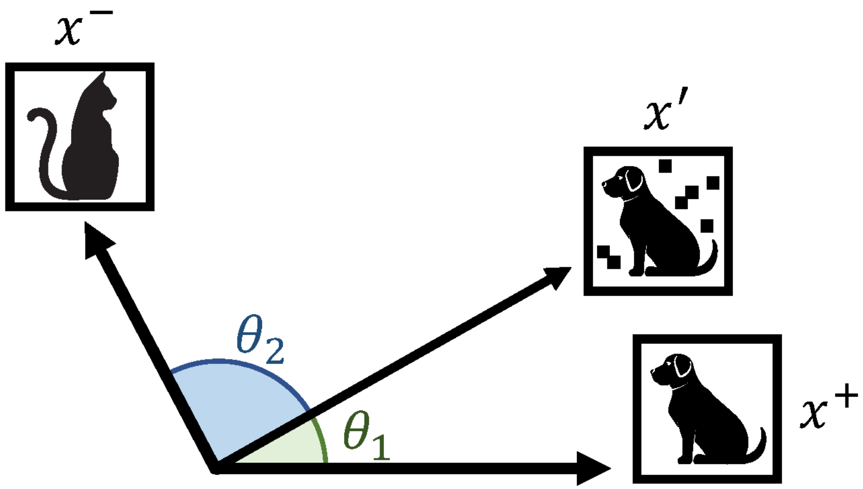
- **Robustness verification** means classifiers whose prediction at point x is verified to be constant within a neighborhood of x, regardless of what a

- Can we design a robustness verification framework for contrastive learning that does not require class labels and downstream tasks?
- Is there any relationship between the robust radius of the CL encoder and that of the downstream task?



RVCL Framework: Verification Problem

- Similar with supervised robustness verification, we define **the conditions under which the disturbance successfully attacks the encoder**.



$$\rho(f(x^+), f(x')) > \rho(f(x^-), f(x'))$$

$$\Updownarrow$$

$$(\tilde{\rho}(f(x^+)) - \tilde{\rho}(f(x^-)))^\top f(x') > 0$$

Definition 4.1 (Verification problem for CL).

$$\widetilde{f}(x^+, x^-, \epsilon) := \min_{x'} \mathbf{W}_{\mathrm{CL}} f(x')$$

$$\text{s.t. } \phi_k(x') = \mathbf{W}_k \widehat{\phi}_{k-1}(x') + \boldsymbol{b}_k, k \in [L],$$

$$\widehat{\phi}_k(x') = \sigma(\phi_k(x')), k \in [L-1],$$

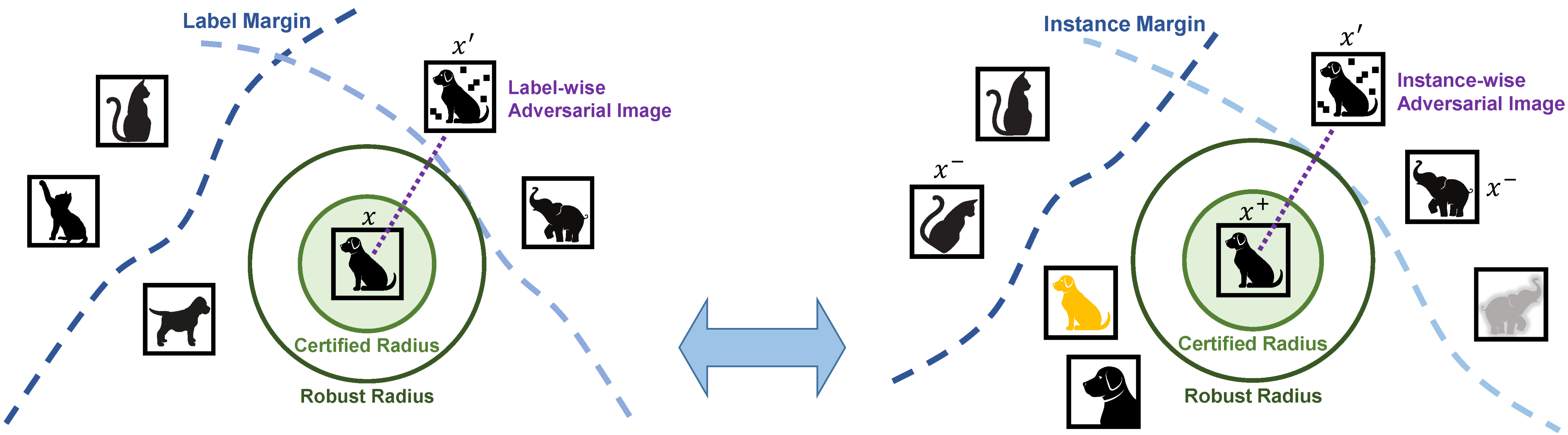
$$\mathbf{W}_{\mathrm{CL}} = \left(\widetilde{\rho}(f(x^+)) - \widetilde{\rho}(f(x^-))\right)^\top \in \mathbb{R}^{1 \times d_L},$$

$$f(x') = \phi_L(x'),\ x' \in \mathcal{B}_\infty(x^+, \epsilon).$$

RVCL Framework: Verification Problem

Supervised

Contrastive



Supervised

$$\widetilde{g}(x, y, \epsilon) := \min_{x'} \; y \cdot g(x')$$

$$\text{s.t. } \phi_k(x') = \mathbf{W}_k \widehat{\phi}_{k-1}(x') + \boldsymbol{b}_k, k \in [L],$$

$$\widehat{\phi}_k(x') = \sigma(\phi_k(x')), k \in [L-1],$$

$$g(x') = \mathbf{W}_{\mathrm{LE}} \phi_L(x') + \boldsymbol{b}_{\mathrm{LE}},$$

$$x' \in \mathcal{B}_\infty(x, \epsilon).$$

Contrastive

$$\widetilde{f}(x^+, x^-, \epsilon) := \min_{x'} \; \mathbf{W}_{\mathrm{CL}} f(x')$$

$$\text{s.t. } \phi_k(x') = \mathbf{W}_k \widehat{\phi}_{k-1}(x') + \boldsymbol{b}_k, k \in [L],$$

$$\widehat{\phi}_k(x') = \sigma(\phi_k(x')), k \in [L-1],$$

$$\mathbf{W}_{\mathrm{CL}} = \left(\widetilde{\rho}(f(x^+)) - \widetilde{\rho}(f(x^-))\right)^\top \in \mathbb{R}^{1 \times d_L},$$

$$f(x') = \phi_L(x'), \; x' \in \mathcal{B}_\infty(x^+, \epsilon).$$

RVCL Framework: Metrics

- By defining the robust radius and certified radius for contrastive learning, we can provide several **robustness metrics** similar to the supervised situation

Robust radius:

$$\mathrm{R}_{\mathrm{CL}}(f; x^+, x^-) := \inf_{\substack{\rho(f(x'), f(x^+)) \\ < \rho(f(x'), f(x^-))}} \|x' - x^+\|_\infty$$

$$= \sup_{\epsilon} \epsilon \text{ s.t. } \widetilde{f}(x^+, x^-, \epsilon) > 0$$

$$\underline{\mathrm{R}}_{\mathrm{CL}}(f; x^+, x^-) := \sup_{\epsilon} \epsilon \text{ s.t. } \underline{f}(x^+, x^-, \epsilon) > 0$$

Average certified radius (ACR) for CL:

$$\mathrm{ACR}_{\mathrm{CL}} := \frac{1}{K|U_{\mathtt{test}}|} \sum_{z \in U_{\mathtt{test}}} \sum_{i=1}^{K} \underline{\mathrm{R}}_{\mathrm{CL}}(f; x^+, x_i^-)$$

Robust instance accuracy:

$$\mathcal{A}_{\mathrm{CL}}^{\epsilon} := \frac{1}{|U_{\mathtt{test}}|} \sum_{z \in U_{\mathtt{test}}} 1_{[\rho(f(x'), f(x^+)) - \rho(f(x'), f(x^-)) > 0]}$$

Certified instance accuracy:

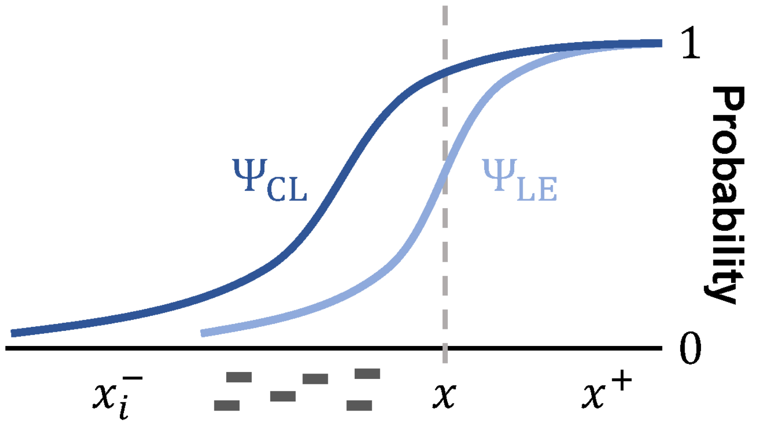
$$\underline{\mathcal{A}}_{\mathrm{CL}}^{\epsilon} := \frac{1}{|U_{\mathtt{test}}|} \sum_{z \in U_{\mathtt{test}}} 1_{[\underline{f}(x^+, x^-, \epsilon) > 0]}$$

RVCL Framework: Theoretical Analysis

- Single positive sample and multiple negative samples:

Theorem 5.3 (Robust radius bound). *Given an encoder $f : \mathcal{X} \to \mathbb{R}^d$ and an unlabeled sample $z = (x^+, \{x_i^-\}_{i=1}^K)$, the downstream predictor $g : \mathbb{R}^d \to \mathbb{R}$ is trained on $\widehat{S} = \{(f(x^+), y_{c^+}), (f(x_i^-), y_{c^-})_{i=1}^K\}$. Then, for different negative samples x_i^-, we have*

$$\mathrm{R}_{\mathrm{CL}}(f; x^+, x_i^-) \geq \mathrm{R}_{\mathrm{LE}}(g; x^+, y_{c^+}).$$



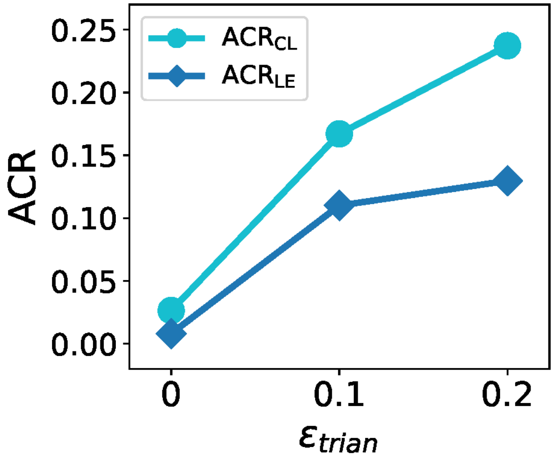
- Multiple positive samples:

Theorem 5.5. *Given an encoder $f : \mathcal{X} \to \mathbb{R}^d$, two positive samples x_1^+, x_2^+ and one negative sample x^-, if $\rho(f(x_1^+), f(x^-)) \geq \rho(f(x_2^+), f(x^-))$, then*

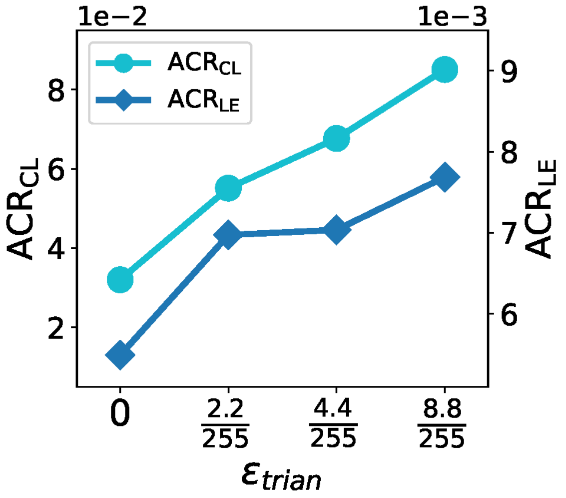
$$\mathrm{R}_{\mathrm{CL}}(f; x_1^+, x^-) \leq \mathrm{R}_{\mathrm{CL}}(f; x_2^+, x^-).$$

Experiments: Average Certified Radius

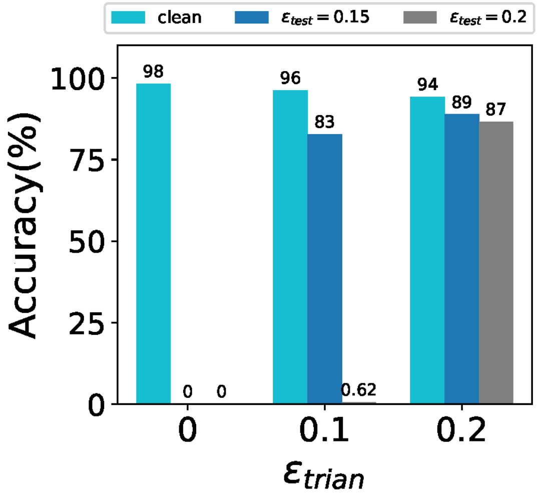
- It is effective to measure the robustness using $\mathrm{ACR}_{\mathrm{CL}}$ without labels and downstream tasks
- $\mathrm{ACR}_{\mathrm{CL}}$ is larger than $\mathrm{ACR}_{\mathrm{LE}}$ with the same ϵ_{train}



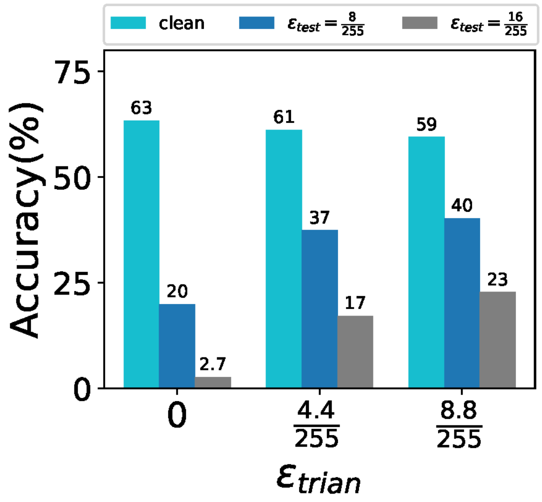
(a) ACR for MNIST



(b) ACR for CIFAR-10



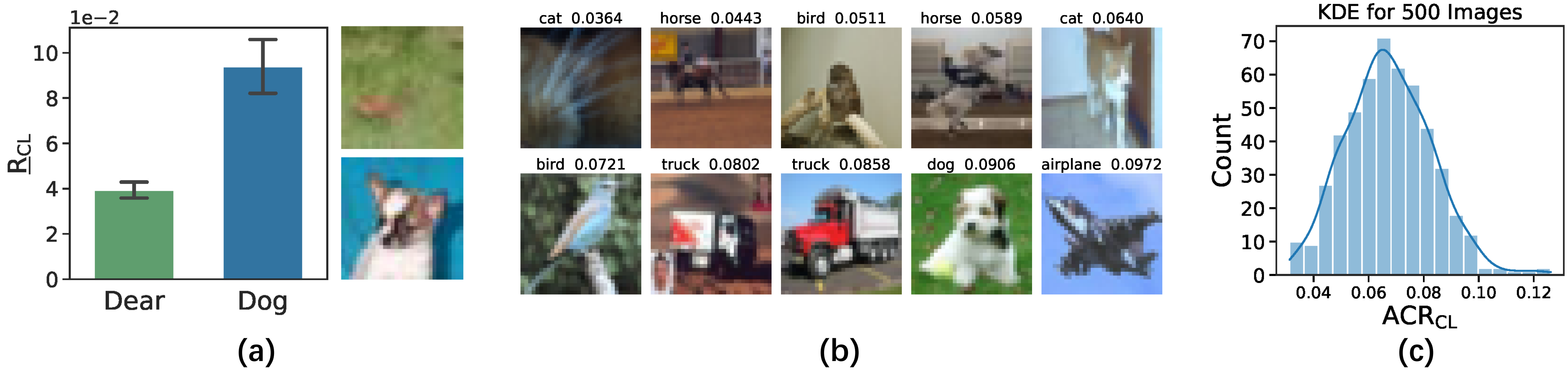
(c) MNIST Robust Test



(d) CIFAR-10 Robust Test

Experiments: Anti-disturbance Ability of Images

- The vague image which is difficult to identify the latent class has a low $\mathrm{ACR}_{\mathrm{CL}}$
- These results verify that $\mathrm{ACR}_{\mathrm{CL}}$ is able to quantify the anti-disturbance ability of images

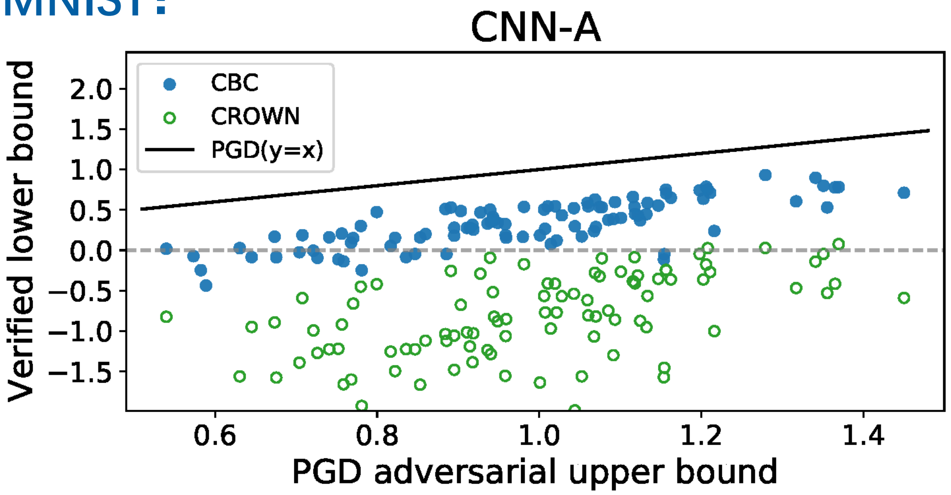


Experiments: Tightness of Verification

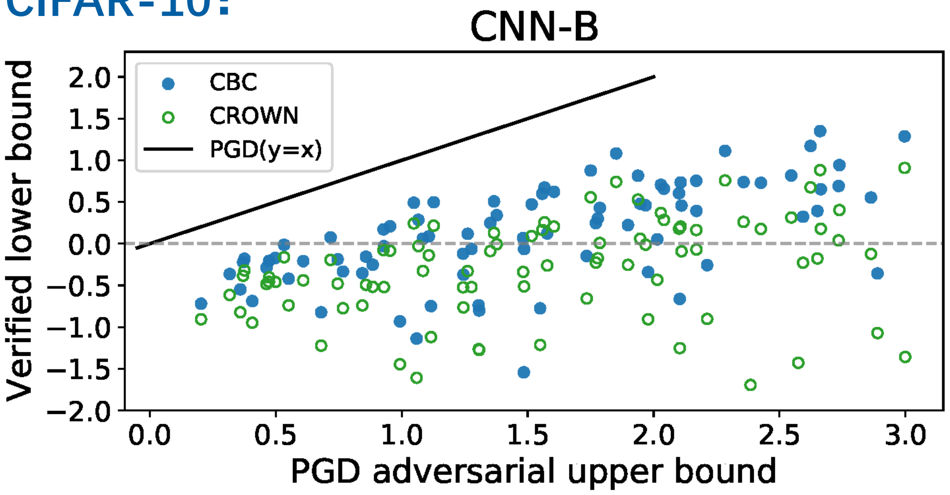
- A stronger supervised verifier can still achieve a tighter certified radius in the RVCL framework

ϵ_{test}	**Model**	ϵ_{train}	Instance Accuracy	Certified Instance Accuracy	
			PGD	**CBC**	**CROWN**
$\frac{2}{255}$	CNN-B	0	100%	97%	96%
		$\frac{2.2}{255}$	100%	100%	100%
$\frac{4}{255}$			91%	26%	11%
		$\frac{4.4}{255}$	100%	55%	34%
		$\frac{8.8}{255}$	100%	68%	52%
	Based		100%	99%	95%
	Deep	$\frac{4.4}{255}$	100%	96%	84%
	CNN-A		99%	91%	81%
$\frac{8}{255}$	CNN-B	$\frac{8.8}{255}$	1%	0%	0%

MNIST:

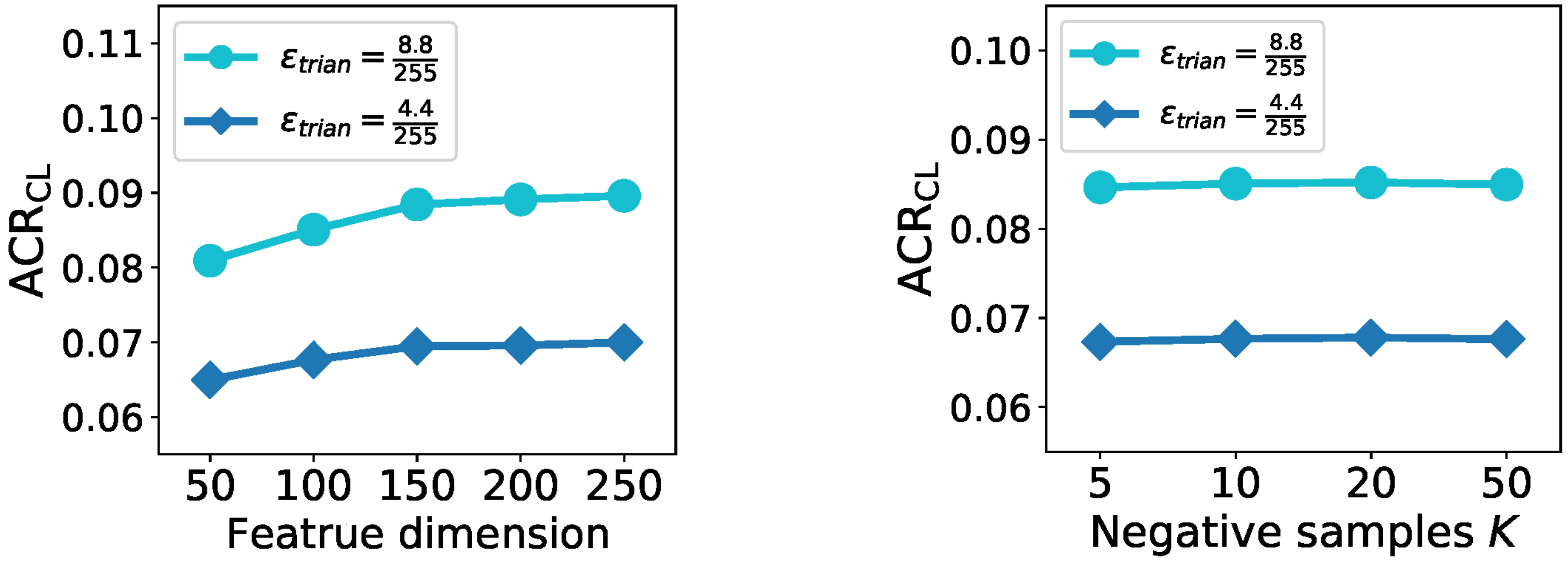


CIFAR-10:



Experiments: Sensitive Analysis

- The results illustrate that $\mathrm{ACR_{CL}}$ is not sensitive to feature dimension and the number of negative samples



THANK YOU