Adapting to Evolving Adversaries with Regularized Continual Robust Training

Sihui Dai*[1,2], Christian Cianfarani*[3], Vikash Sehwag[4], Prateek Mittal[2], Arjun Bhagoji[5]

[1]CapitalOne [2]Department of Electrical and Computer Engineering, Princeton University
[3]Department of Computer Science, University of Chicago [4]Google Deepmind
[5]Center for Machine Intelligence and Data Science, Indian Institute of Technology, Bombay

ICML 2025, Vancouver BC

Adversarial Training

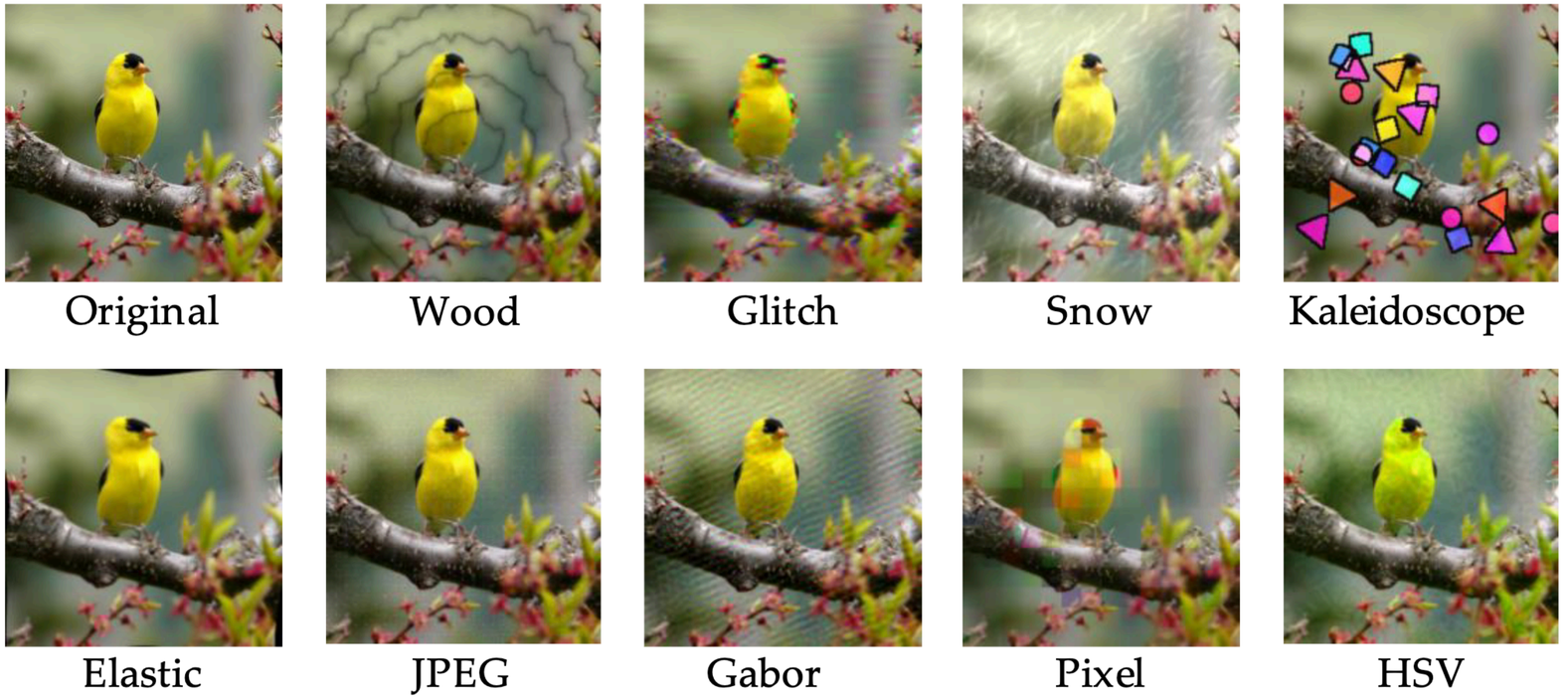
RobustBench Leaderboards Paper FAQ Contribute Model Zoo 🚀

Leaderboard: CIFAR-10, $\ell_\infty = 8/255$, untargeted attack

Show 15 entries Search: Papers, architectures, ve

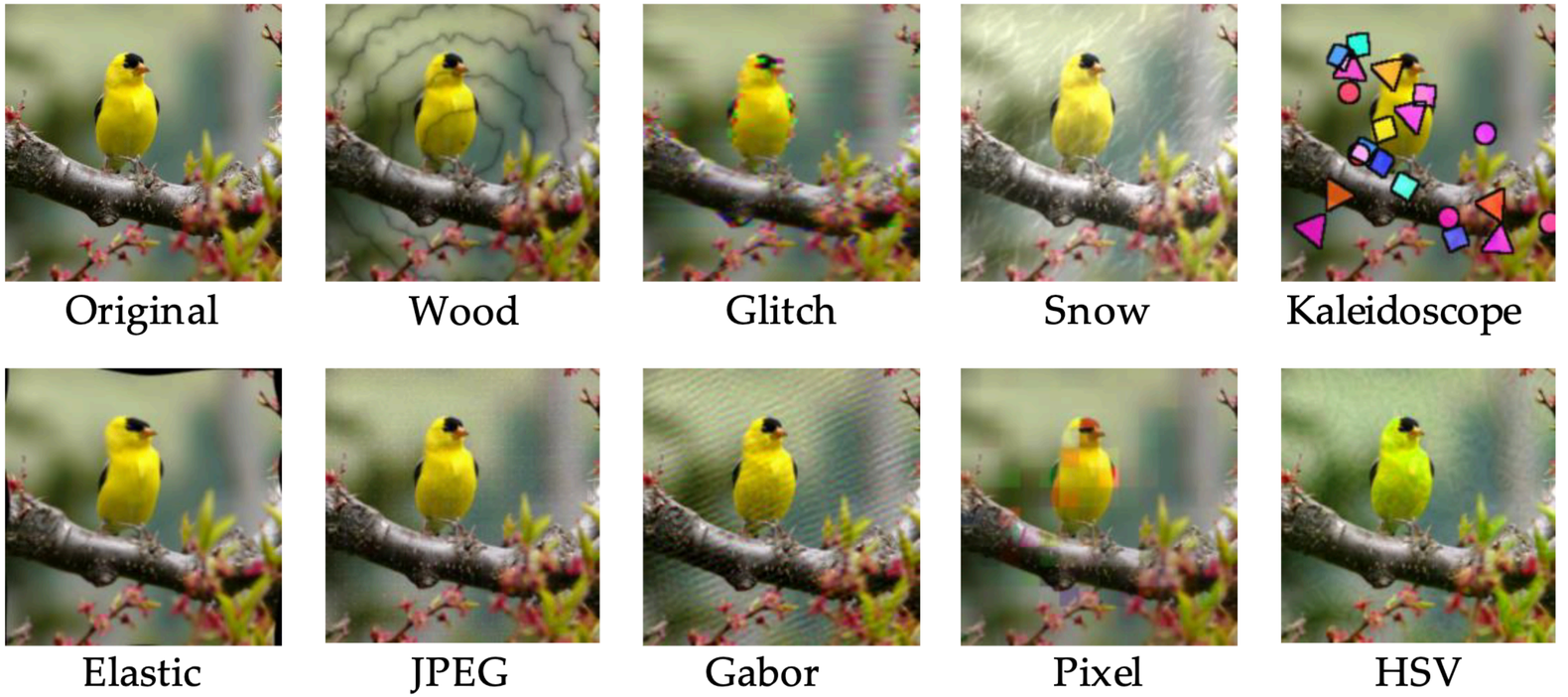
Rank	Method	Standard accuracy	AutoAttack robust accuracy	Best known robust accuracy	AA eval. potentially unreliable	Extra data	Architecture	Venue
1	Adversarial Robustness Limits via Scaling-Law and Human-Alignment Studies *It uses additional 300M synthetic images in training.*	93.68%	73.71%	73.71%	✗	✗	WideResNet-94-16	ICML 2024
2	MeanSparse: Post-Training Robustness Enhancement Through Mean-Centered Feature Sparsification *It adds the MeanSparse operator to the adversarially trained models.*	93.24%	72.08%	72.08%	✗	☑	MeanSparse RaWideResNet-70-16	arXiv, Jun 2024
3	Adversarial Robustness Limits via Scaling-Law and Human-Alignment Studies *It uses additional 300M synthetic images in training.*	93.11%	71.59%	71.59%	✗	✗	WideResNet-82-8	ICML 2024
4	Robust Principles: Architectural Design Principles for Adversarially Robust CNNs *It uses additional 50M synthetic images in training.*	93.27%	71.07%	71.07%	✗	✗	RaWideResNet-70-16	BMVC 2023
5	Better Diffusion Models Further Improve Adversarial Training *It uses additional 50M synthetic images in training.*	93.25%	70.69%	70.69%	✗	✗	WideResNet-70-16	ICML 2023

Types of Adversaries



Kaufmann, Maximilian, et al. *Evaluating Robustness to Unforeseen Adversarial Attacks*. 2023.

Types of Adversaries



Can we train models to be robust against multiple adversaries?

Kaufmann, Maximilian, et al. *Evaluating Robustness to Unforeseen Adversarial Attacks*. 2023.

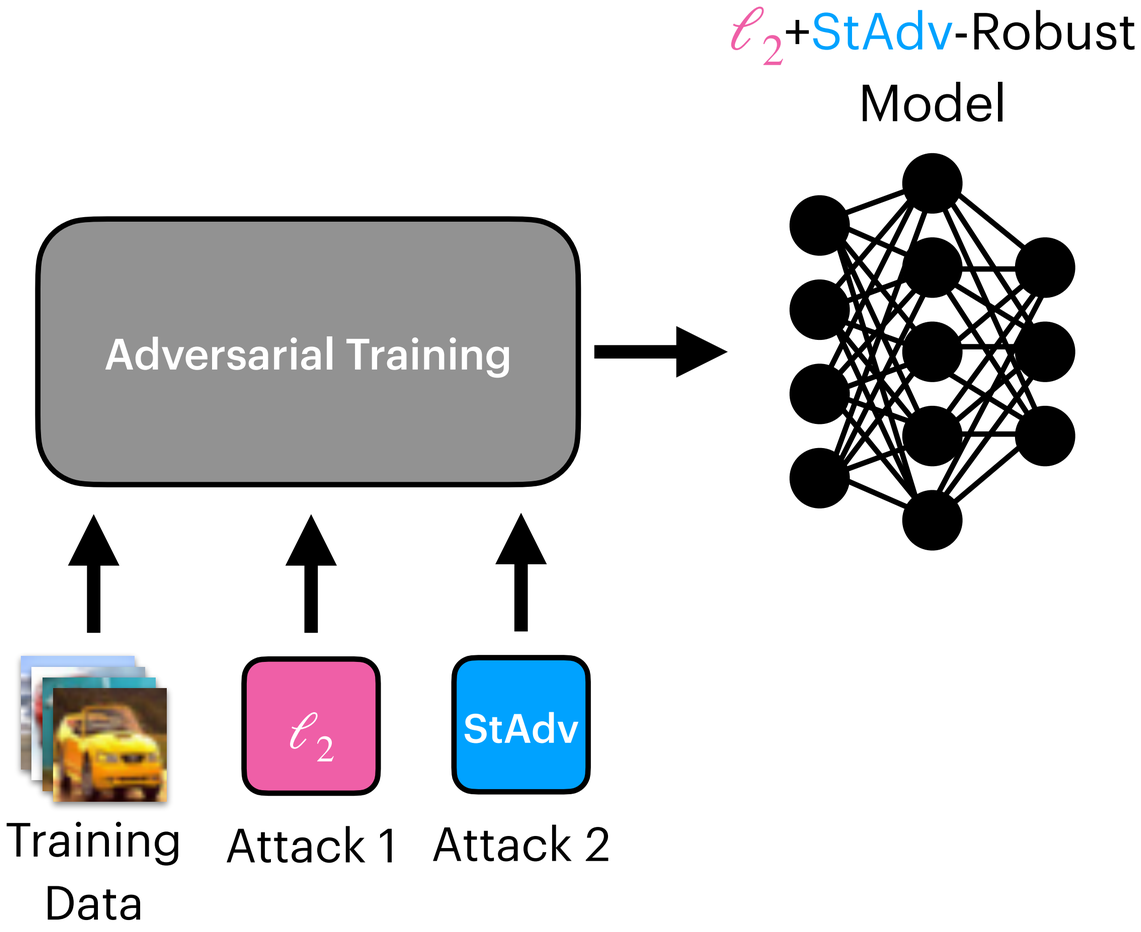
Types of Adversaries

Can we train models to be robust against multiple adversaries?

Can we *update* existing models in the face of new adversaries?

Kaufmann, Maximilian, et al. *Evaluating Robustness to Unforeseen Adversarial Attacks*. 2023.

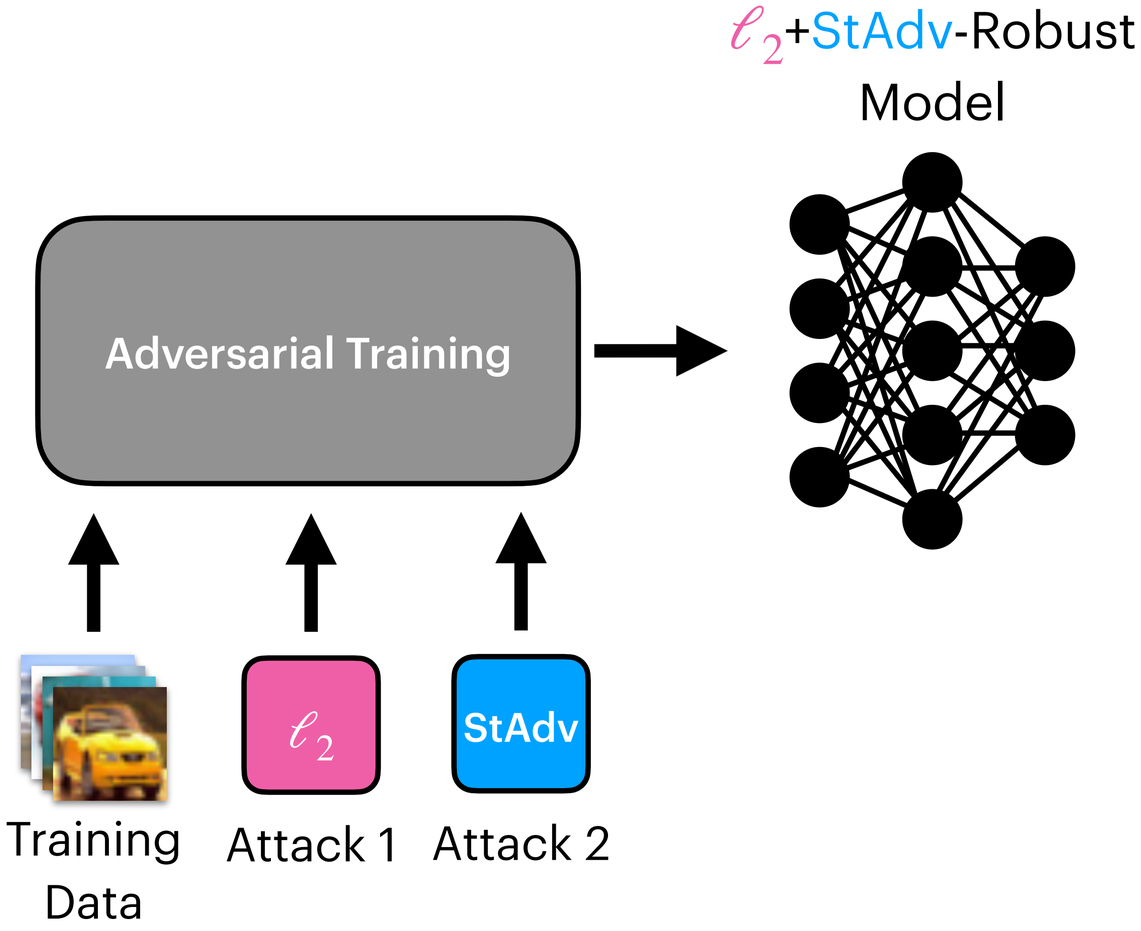
Training Multi-Robust Models from Scratch



[1] Maini et al. *Adversarial robustness against the union of multiple perturbation models*. ICML 2020
[2] Tramèr and Boneh. *Adversarial Training and Robustness for Multiple Perturbations*. Neurips 2019
[3] Madaan et al. *Learning to generate noise for robustness against multiple perturbations*. ICML 2021
[4] Croce and Hein. *Provable robustness against all adversarial ℓ_p-perturbations for $p \geq 1$*. ICLR 2020
[5] Jiang and Singh. *Ramp: Boosting adversarial robustness against multiple ℓ_p perturbations for universal robustness*. Neurips 2024

Training Multi-Robust Models from Scratch

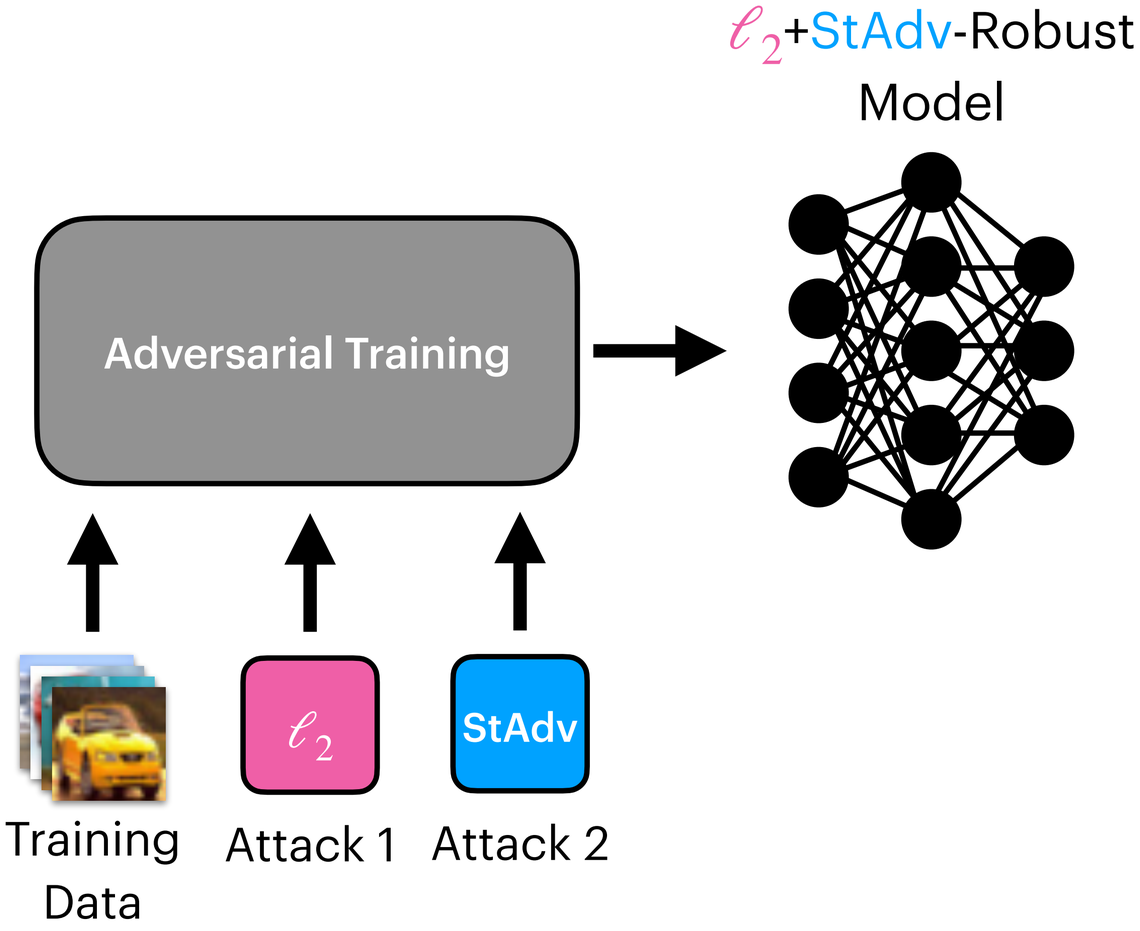
- Multiple existing techniques can train models robust to multiple attacks [1-5].



[1] Maini et al. *Adversarial robustness against the union of multiple perturbation models*. ICML 2020
[2] Tramèr and Boneh. *Adversarial Training and Robustness for Multiple Perturbations*. Neurips 2019
[3] Madaan et al. *Learning to generate noise for robustness against multiple perturbations*. ICML 2021
[4] Croce and Hein. *Provable robustness against all adversarial ℓ_p-perturbations for $p \geq 1$*. ICLR 2020
[5] Jiang and Singh. *Ramp: Boosting adversarial robustness against multiple ℓ_p perturbations for universal robustness*. Neurips 2024

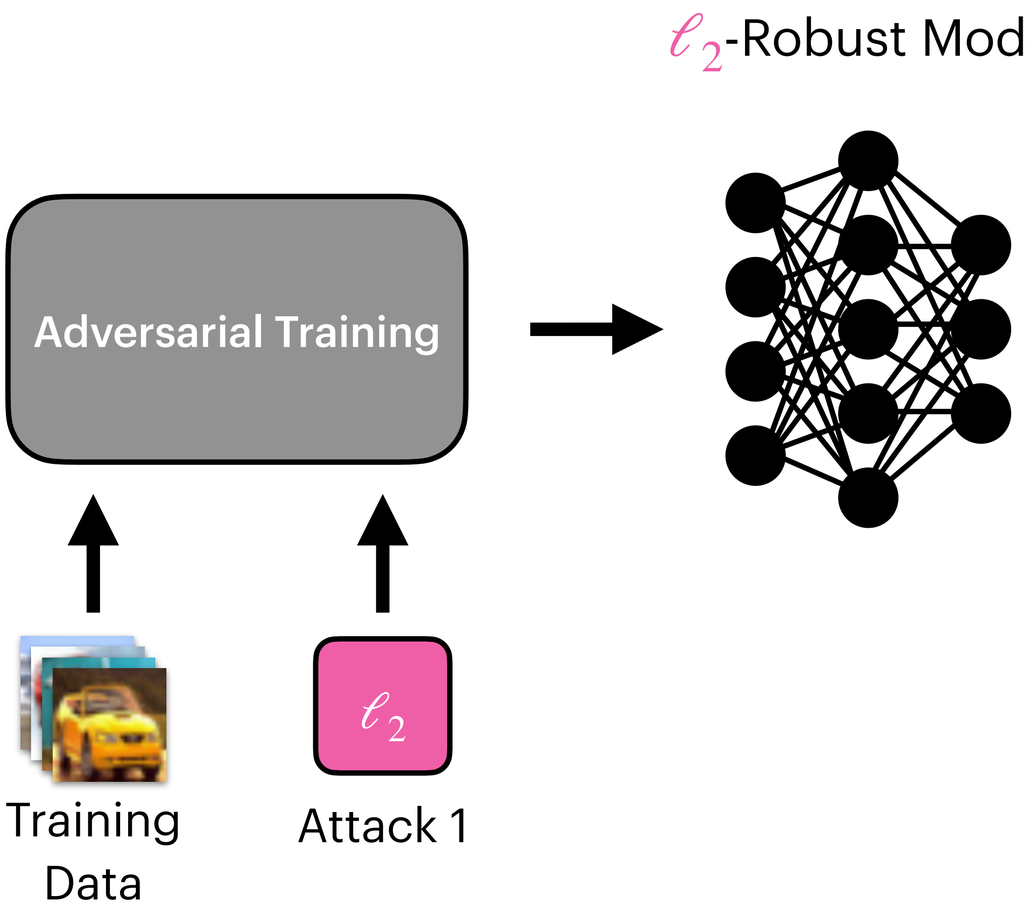
Training Multi-Robust Models from Scratch

- Multiple existing techniques can train models robust to multiple attacks [1-5].
- If we become aware of a new attack after training, we have to retrain from scratch!

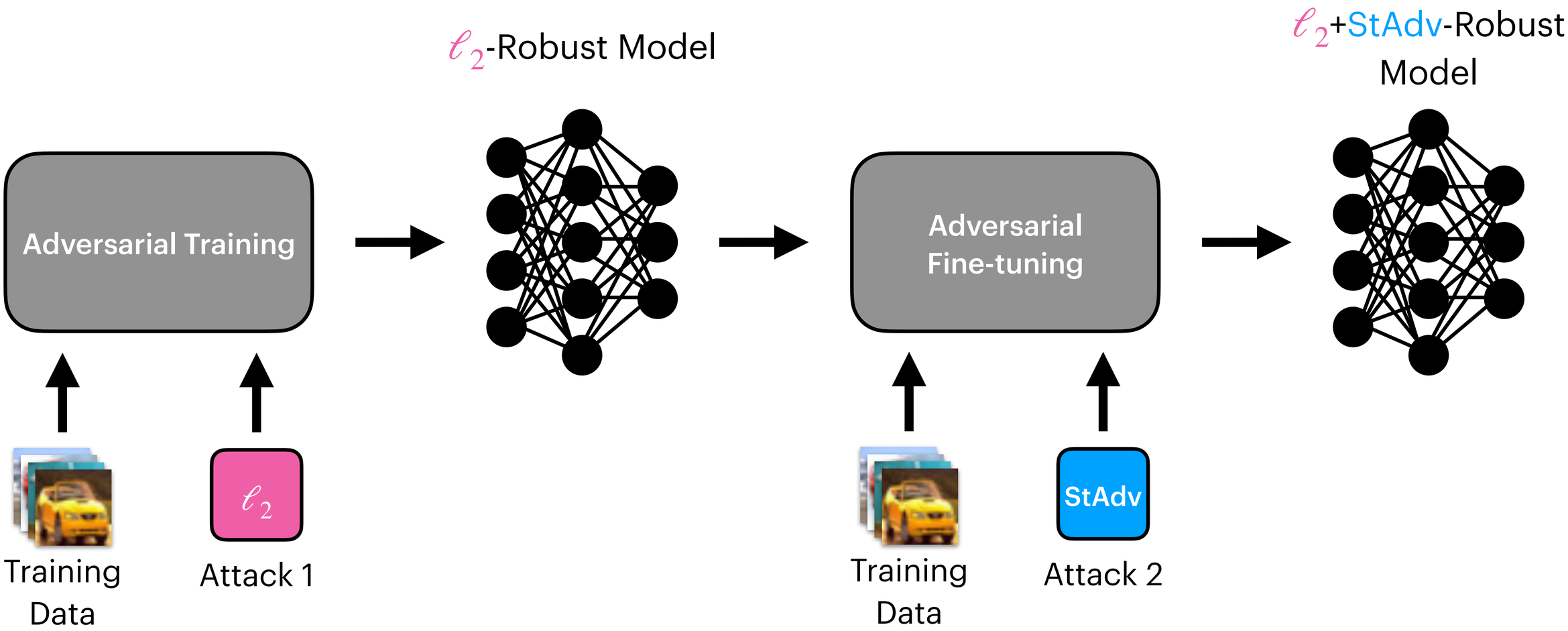


[1] Maini et al. *Adversarial robustness against the union of multiple perturbation models*. ICML 2020
[2] Tramèr and Boneh. *Adversarial Training and Robustness for Multiple Perturbations*. Neurips 2019
[3] Madaan et al. *Learning to generate noise for robustness against multiple perturbations*. ICML 2021
[4] Croce and Hein. *Provable robustness against all adversarial ℓ_p-perturbations for $p \geq 1$*. ICLR 2020
[5] Jiang and Singh. *Ramp: Boosting adversarial robustness against multiple ℓ_p perturbations for universal robustness*. Neurips 2024

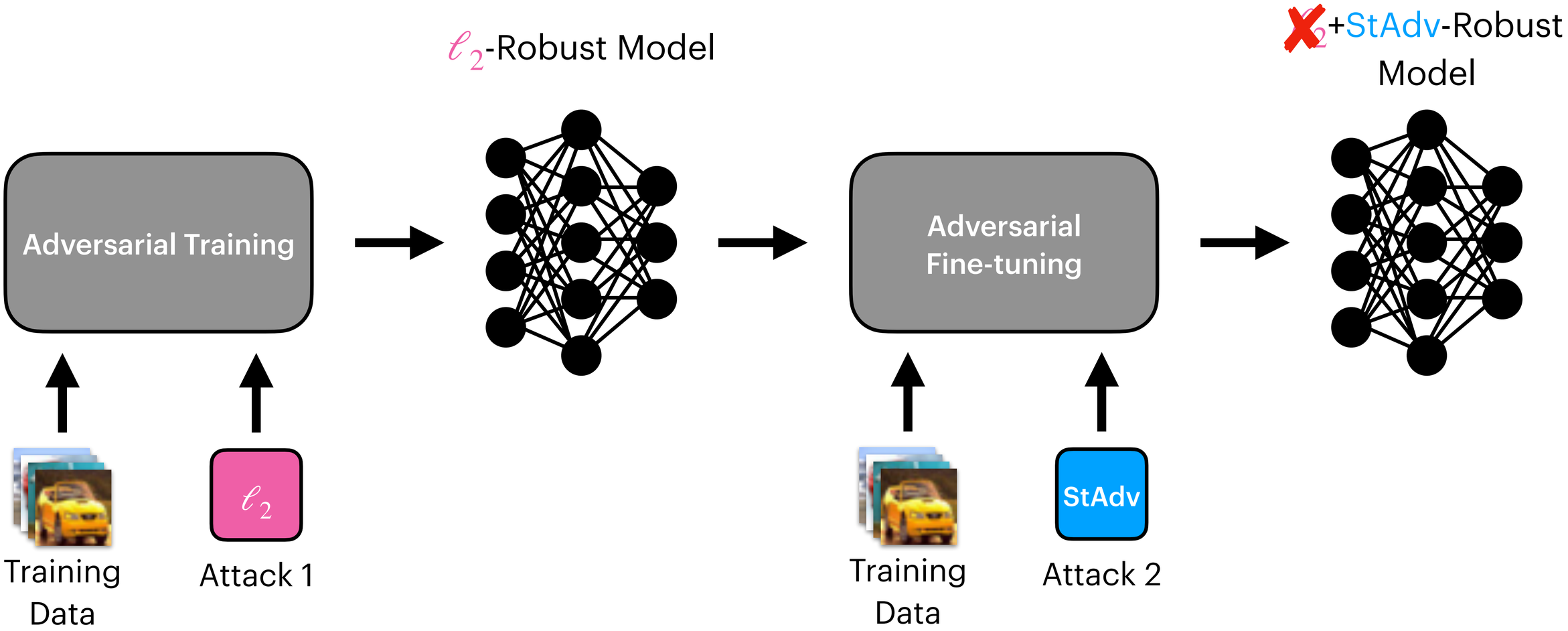
Idea: Fine-tune on New Attack



Idea: Fine-tune on New Attack

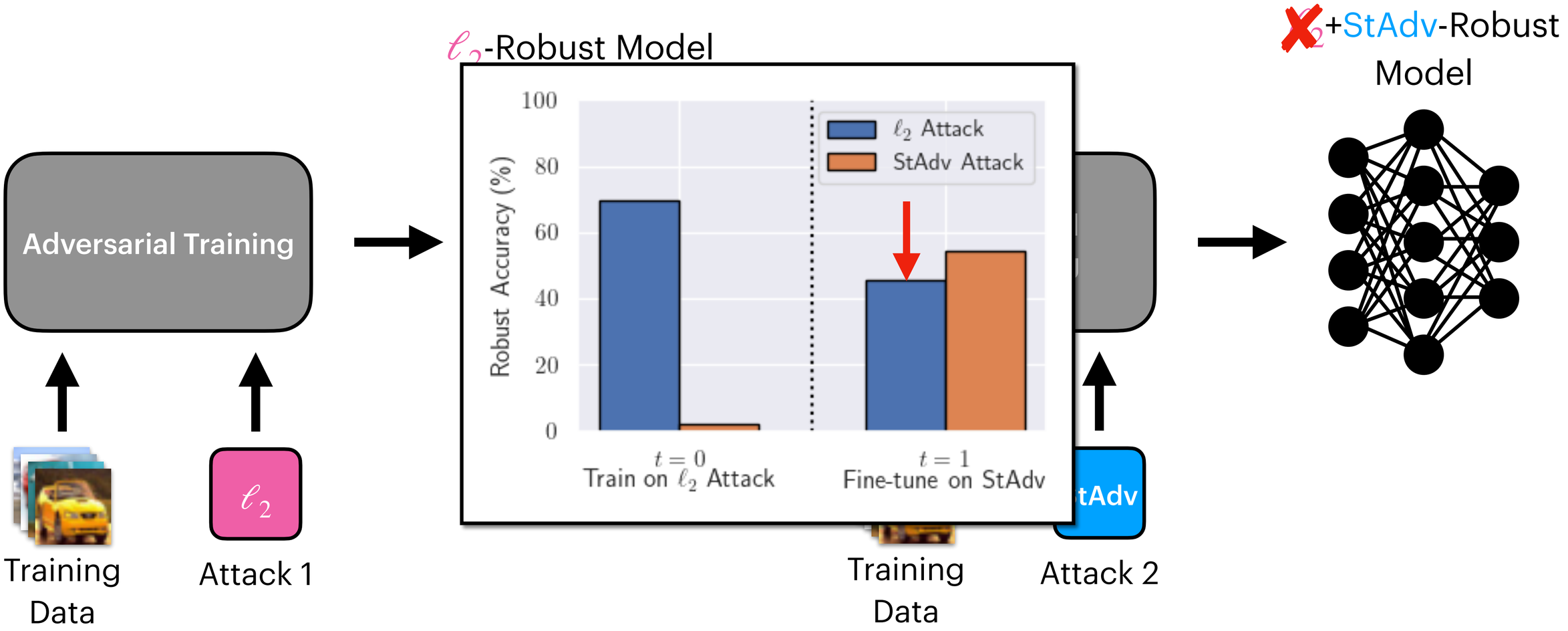


Idea: Fine-tune on New Attack



Problem: Fine-tuned model loses robustness against first attack!

Idea: Fine-tune on New Attack



Problem: Fine-tuned model loses robustness against first attack!

Improvement #1: Regularization

Improvement #1: Regularization

How can we learn robust representations that easily transfer to new attacks?

Improvement #1: Regularization

How can we learn robust representations that easily transfer to new attacks?

Theorem (Informal)

Let $h : \mathbb{R}^d \to \mathbb{R}^c$, mapping inputs to logits.

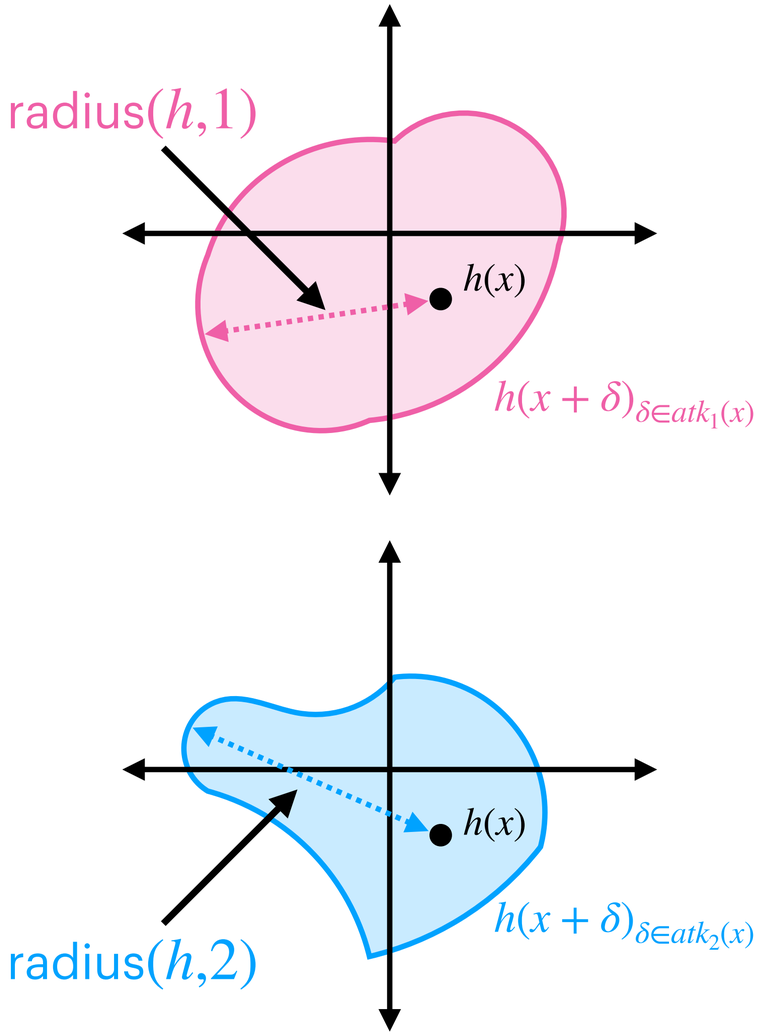
Improvement #1: Regularization

How can we learn robust representations that easily transfer to new attacks?

Theorem (Informal)

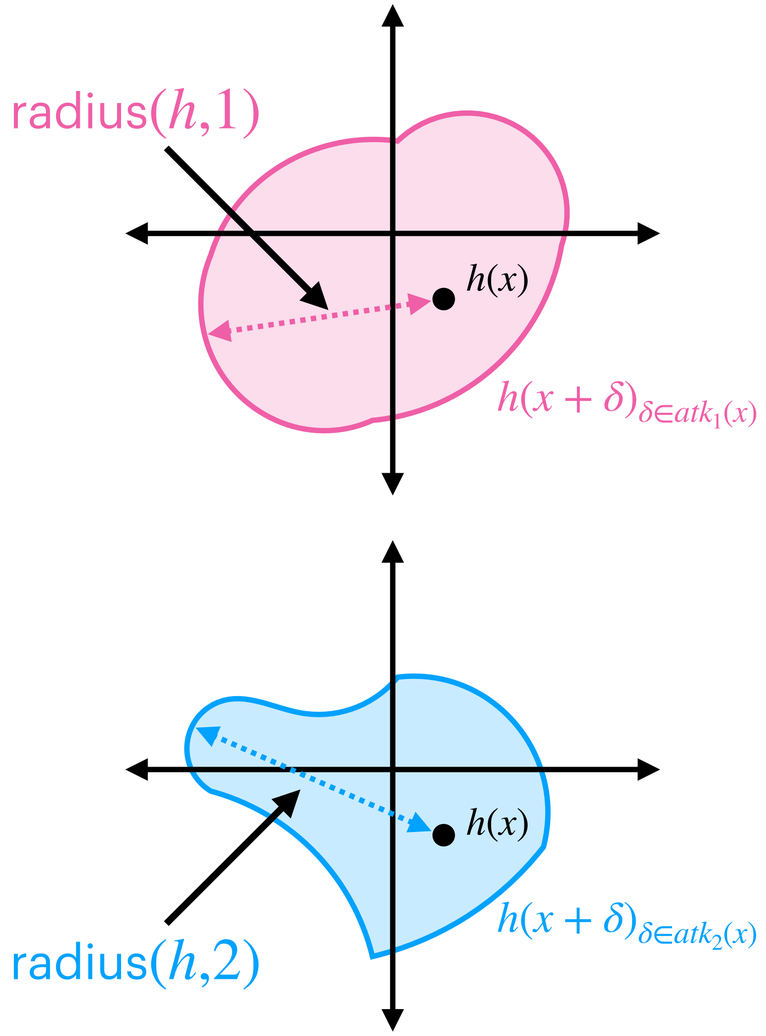
Let $h : \mathbb{R}^d \to \mathbb{R}^c$, mapping inputs to logits.

Let $\text{radius}(h, i) = \mathbb{E}_x[\max_{\delta \in atk_i(x)} \|h(x+\delta) - h(x)\|_2]$.



Improvement #1: Regularization

How can we learn robust representations that easily transfer to new attacks?



Theorem (Informal)

Let $h : \mathbb{R}^d \to \mathbb{R}^c$, mapping inputs to logits.

Let $\text{radius}(h, i) = \mathbb{E}_x[\max_{\delta \in atk_i(x)} \|h(x+\delta) - h(x)\|_2]$.

Let $L_i(h)$ be the adversarial loss of h w.r.t attack i.

$$L_{1,2}(h) - L(h) \leq M \sum_{i \in 1,2} \text{radius}(h, i) + C\,.$$

Proposed Regularization Term

$$L_{\mathrm{ALR}}(h,t) = L(h,t) + \lambda R_{\mathrm{ALR}}(h,t),$$

where

$$R_{\mathrm{ALR}} = \frac{1}{m}\sum_{i=1}^{m} \max_{\delta \in atk(x_i)} \|h(x_i+\delta) - h(x_i)\|_2$$

Improvement #2: Fine-tuning with Replay

[1] Croce, Francesco, and Matthias Hein. *Adversarial robustness against multiple l_p-threat models at the price of one and how to quickly fine-tune robust models to another threat model*. ICML 2022.

Improvement #2: Fine-tuning with Replay

- Sample an attack for each training batch:
  1. Before batch b, compute running error for each attack: $\text{rerr}_{b,atk}$
  2. Sample attack $atk \in atks$ with probability

$$\frac{\text{rerr}_{b,atk}}{\sum_{a \in atks} \text{rerr}_{b,a}}$$

[1] Croce, Francesco, and Matthias Hein. *Adversarial robustness against multiple l_p-threat models at the price of one and how to quickly fine-tune robust models to another threat model*. ICML 2022.

Improvement #2: Fine-tuning with Replay

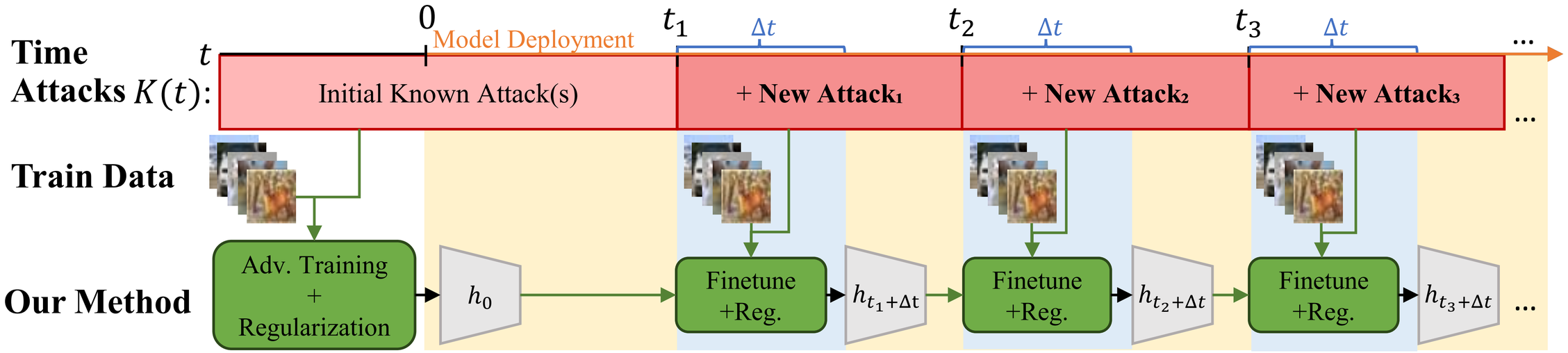
- Sample an attack for each training batch:
  1. Before batch b, compute running error for each attack: $\text{rerr}_{b,atk}$
  2. Sample attack $atk \in atks$ with probability

$$\frac{\text{rerr}_{b,atk}}{\sum_{a \in atks} \text{rerr}_{b,a}}$$

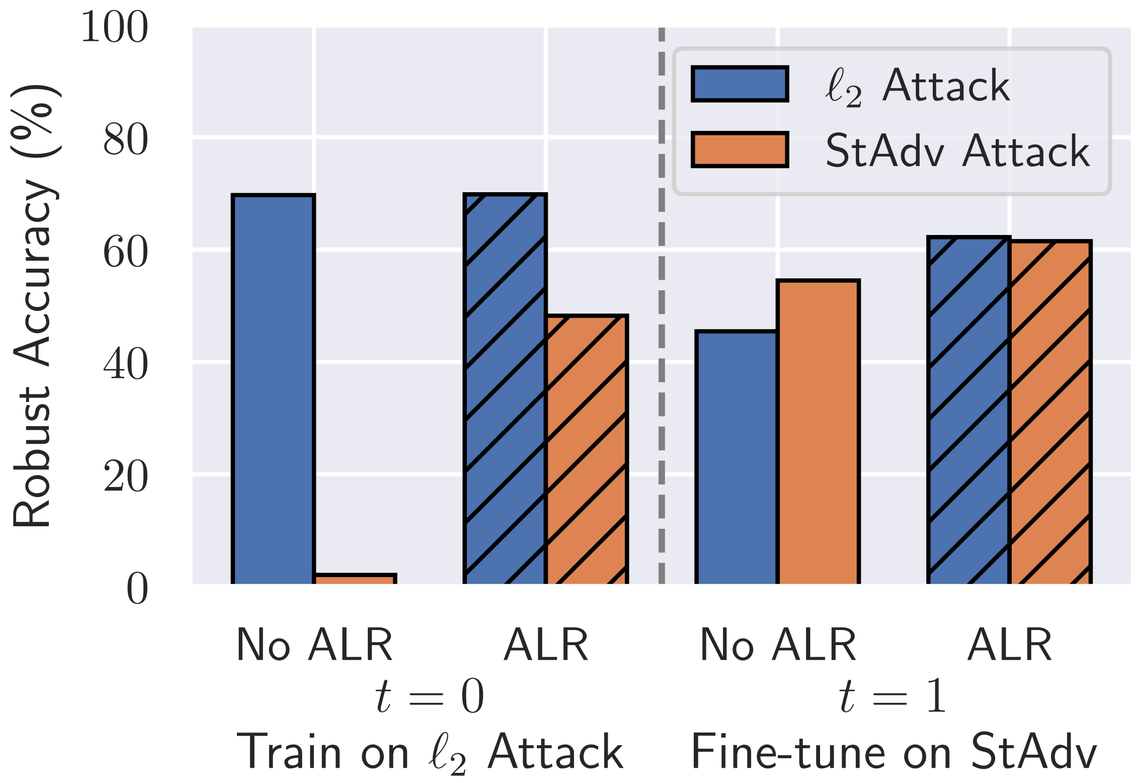
- Originally proposed by Croce and Hein [1] for ℓ_p-bounded attacks
  - We extend this approach to arbitrary adversarial attacks

[1] Croce, Francesco, and Matthias Hein. *Adversarial robustness against multiple l_p-threat models at the price of one and how to quickly fine-tune robust models to another threat model*. ICML 2022.

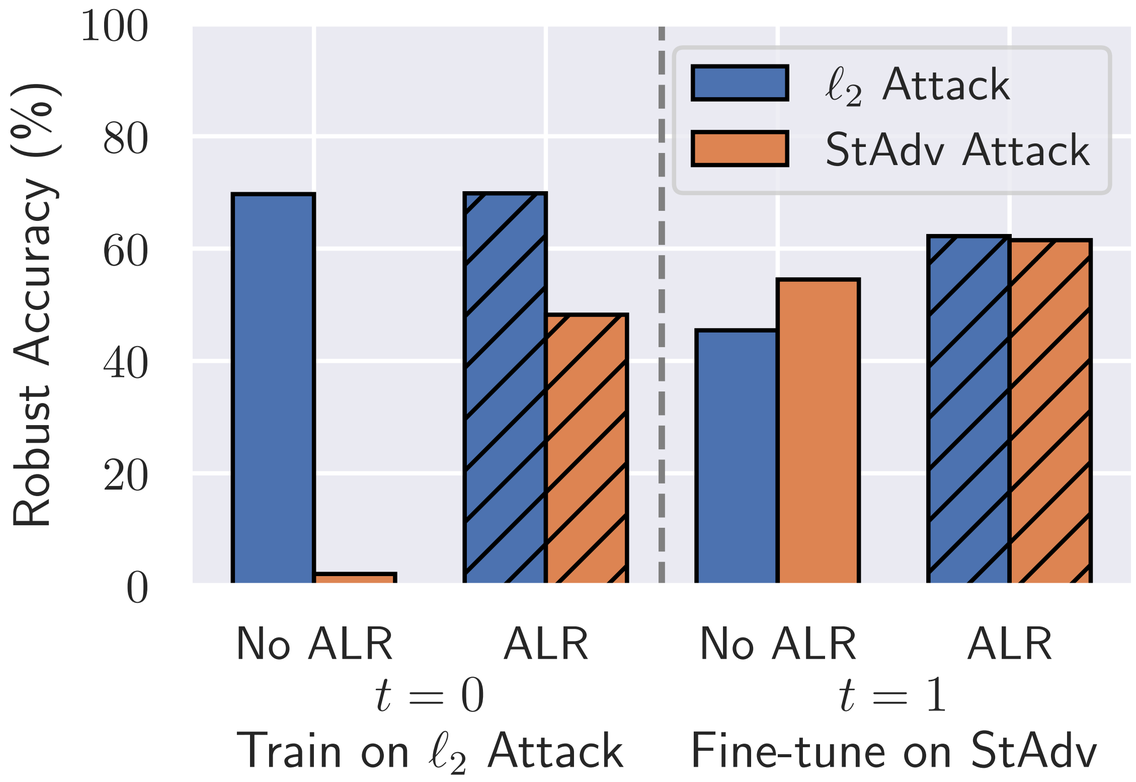
Regularized Continual Robust Training (RCRT)



Improved Robust Accuracy with RCRT

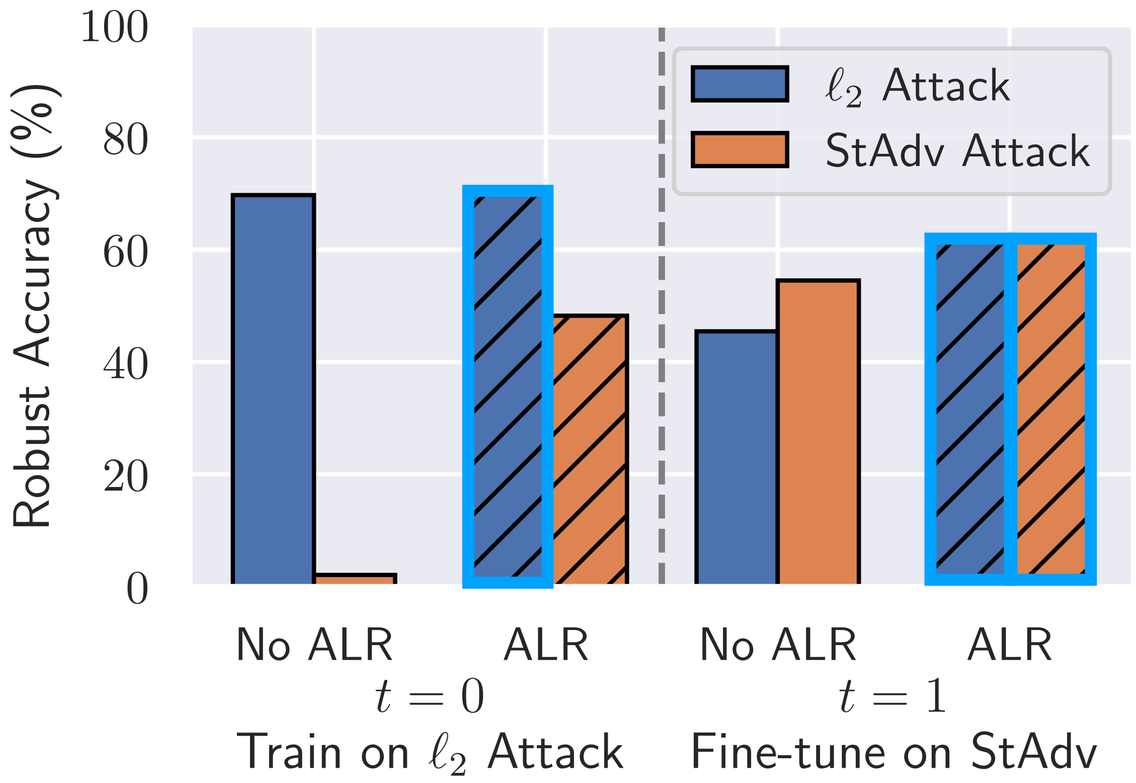


Improved Robust Accuracy with RCRT



Our method:

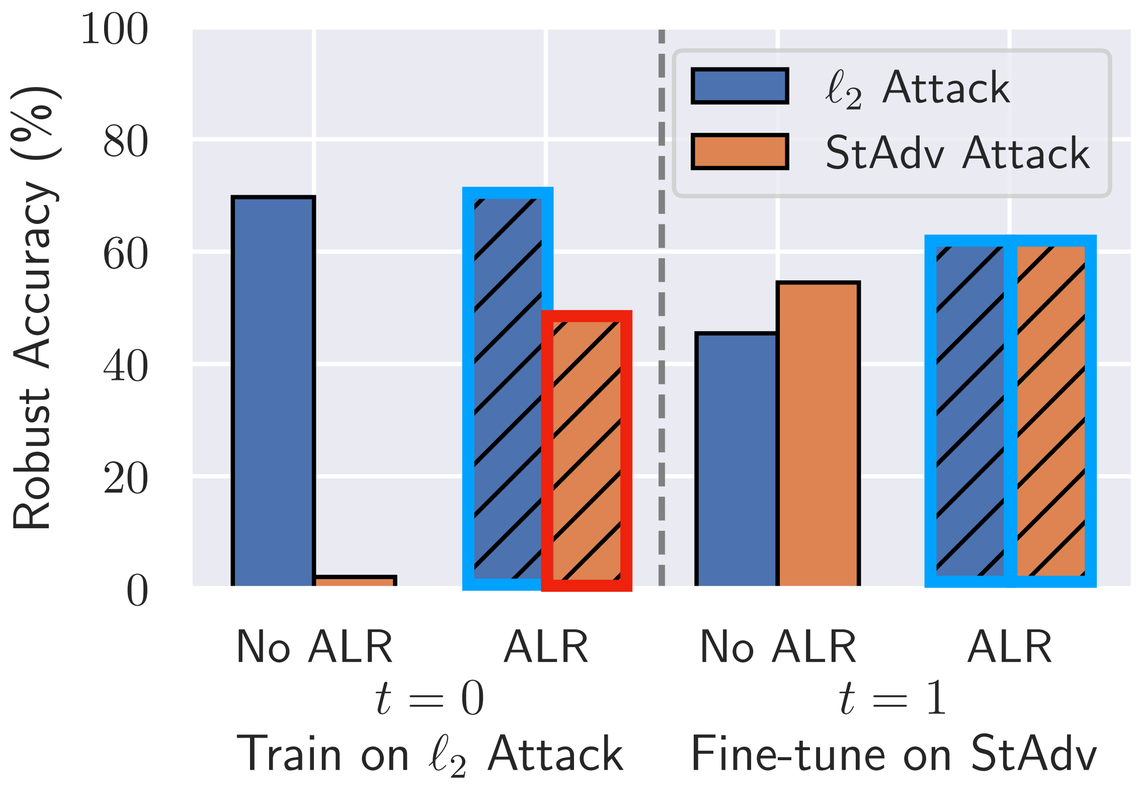
Improved Robust Accuracy with RCRT



Our method:

1. Reduces forgetting of previously seen attacks

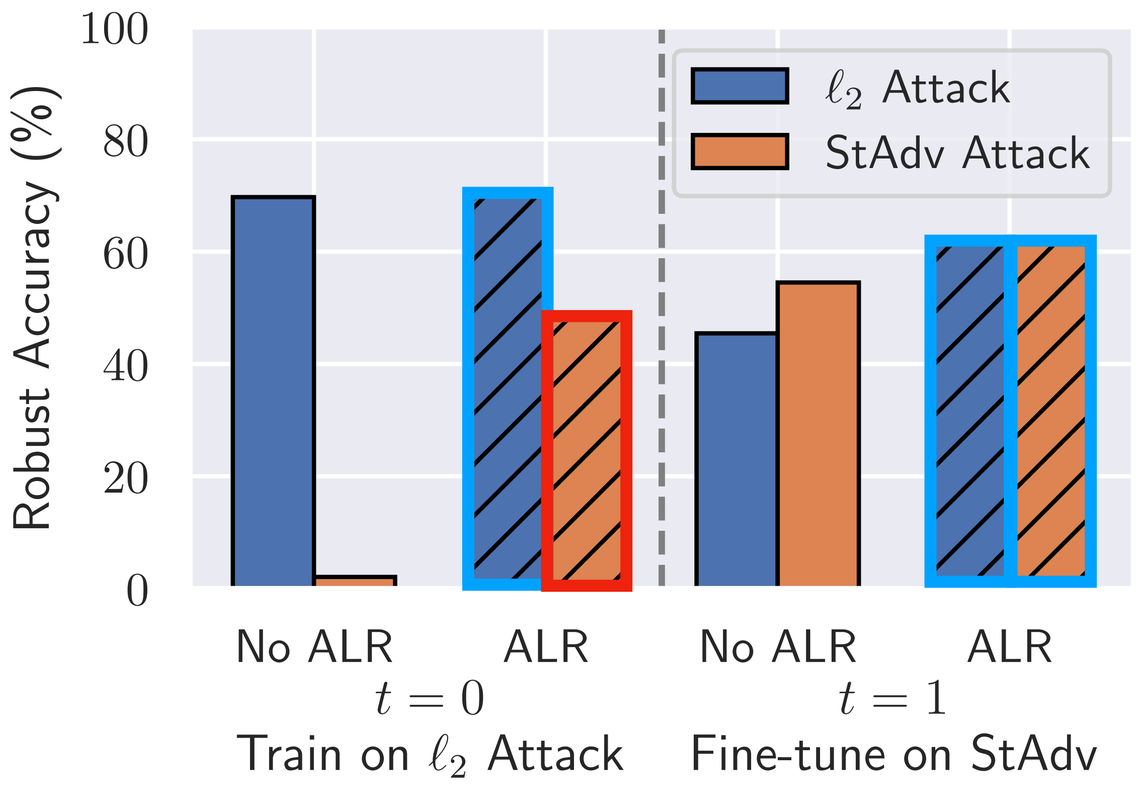
Improved Robust Accuracy with RCRT



Our method:

1. Reduces forgetting of previously seen attacks
2. Performs well on unseen attacks

Improved Robust Accuracy with RCRT



Our method:

1. Reduces forgetting of previously seen attacks
2. Performs well on unseen attacks
3. Saves time over training from scratch (over 3x as fast for each time step)

Conclusions and Future Work

- We demonstrate the use of fine-tuning to efficiently gain robustness against arbitrary attackers
- Our theoretical results point towards simple regularization methods for improving robustness to seen and unseen attacks
- Additional work is necessary to outperform existing baselines in additional settings and to better understand the theoretical limits of multi-attack training

Poster: Tuesday, July 15, 4:30-7:00pm