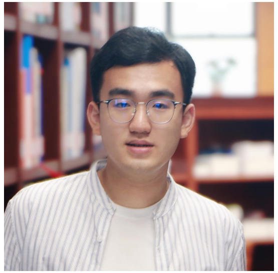
Indiscriminate Poisoning Attacks on Unsupervised Contrastive Learning

Hao He*



Kaiwen Zha*

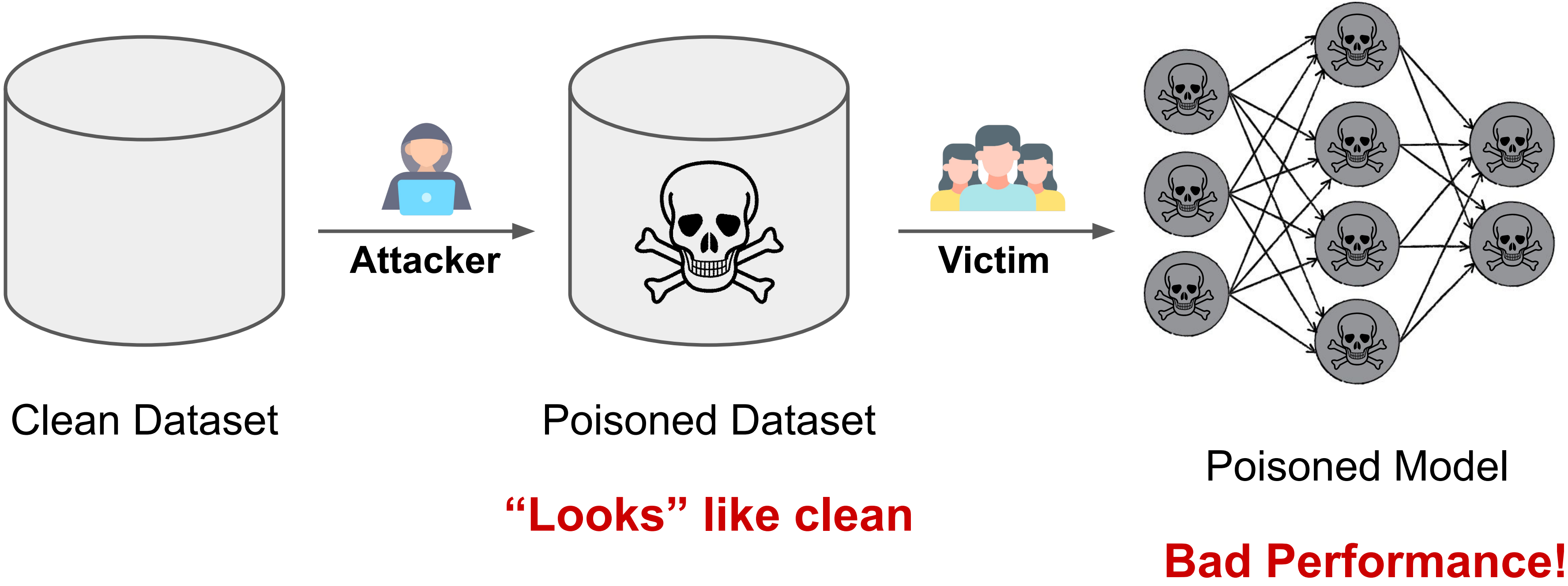
Dina Katabi

(* indicates equal contribution)

ICLR 2023 Spotlight (notable top 25%)

https://kaiwenzha.github.io/contrastive-poisoning

Indiscriminate Data Poisoning



Prior Indiscriminate Poisoning Methods are Successful, but

[1] TensorClog (Shen et al., 2019)

[2] Alignment (Fowl et al., 2021)

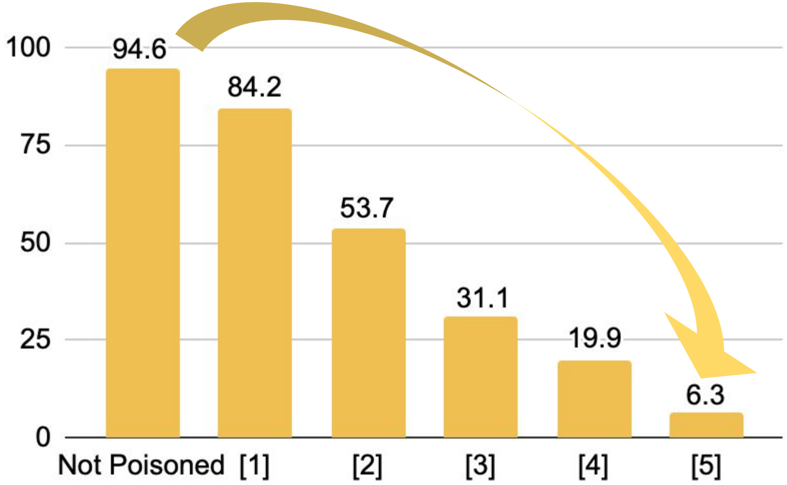
[3] DeepConfuse (Feng et al. 2019)

[4] Unlearnable Example (Huang et al., 2021)

[5] Adversarial Poisoning (Fowl et al., 2021)

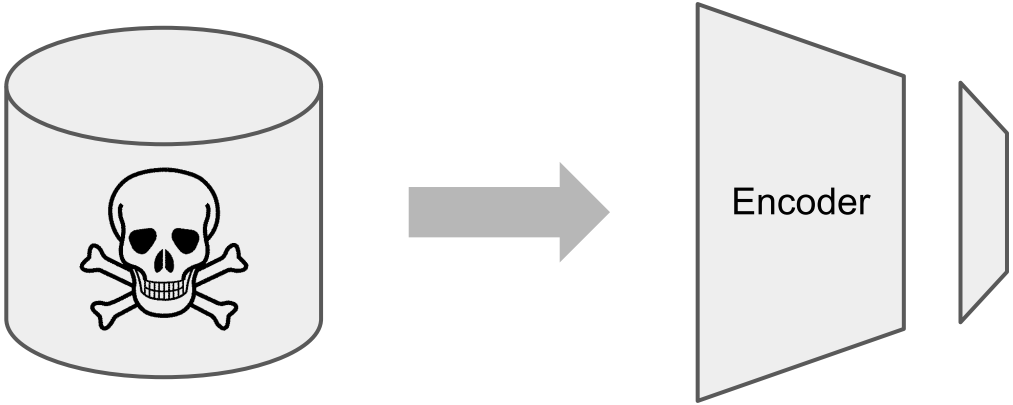
…

Poisoned Accuracy on CIFAR-10



Prior Works Only Attack Supervised Learning!

Contrastive Learning Can Break Prior Attacks

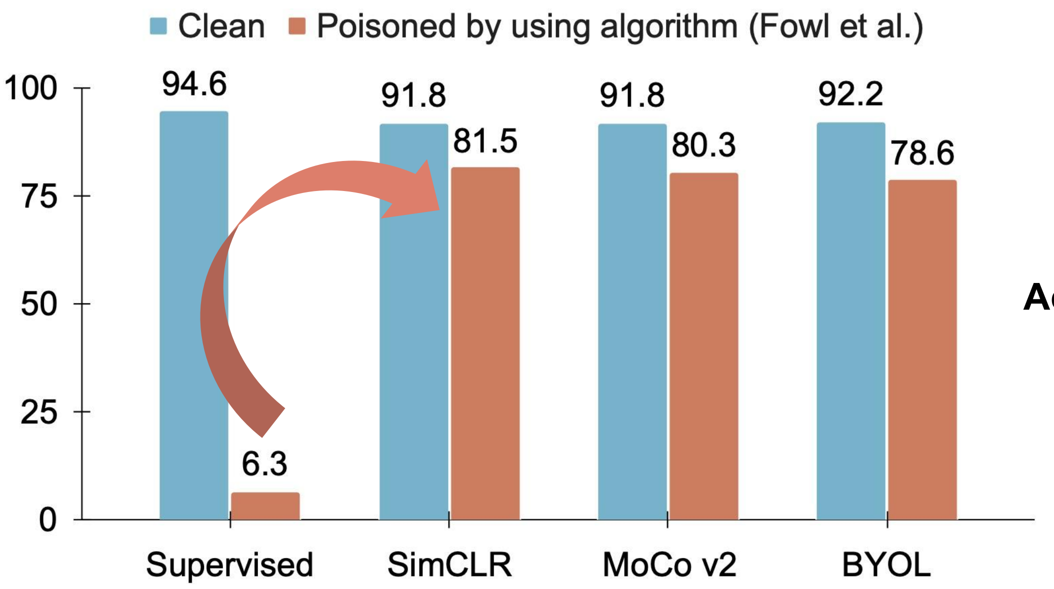


Poisoned Dataset

Step 2: Learning a classifier with the representations fixed

Step 1: Learning representations via contrastive learning

Contrastive Learning Can Break Prior Attacks



Accuracy on CIFAR-10

Our Problem: How to Poison Contrastive Learning?

Our Idea: Shortcutting the Contrastive Learning

Shortcut: providing the model an *easy* way to minimize the contrastive learning loss *without* actually learning real features.

Poisoned model aligns poisoned views via the perturbation

Poisoned model does not align clean views

Contrastive Poisoning (CP)

Optimize the poison to minimize the contrastive learning (CL) loss

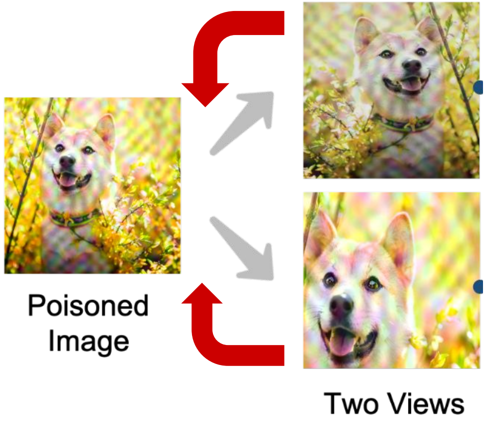
Contrastive Learning Loss

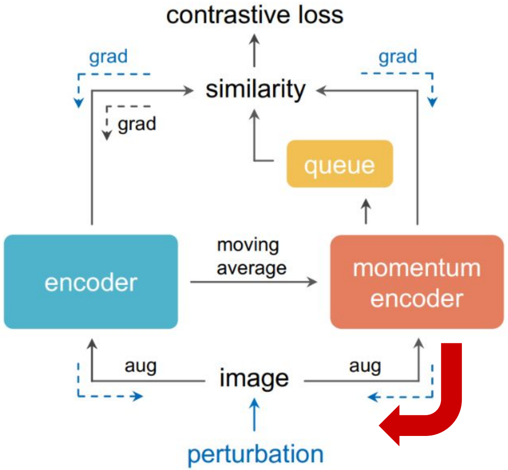
$$\min_{\theta,\delta:\|\delta(x)\|_\infty \le \epsilon} \mathbb{E}_{\{x_i\}_{i=1}^B \sim \mathcal{D}_c} \mathcal{L}_{\mathrm{CL}}(f_\theta; \{x_i + \delta(x_i)\}_{i=1}^B)$$

Poisoning Perturbation

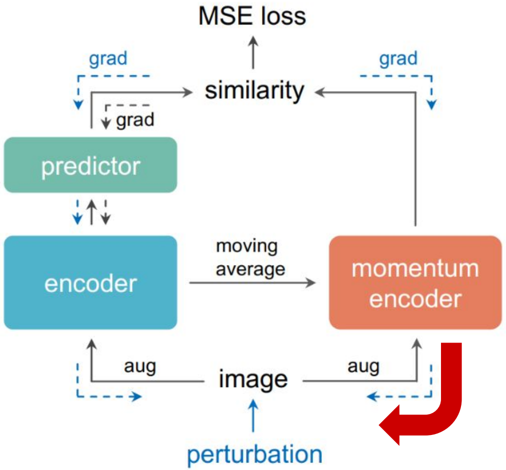
Neural Network

Contrastive Poisoning (CP)





MoCo v2



BYOL

Back-propagate through
Data Augmentation

Back-propagate through
Momentum Encoder

Results - Same Contrastive Learning Algorithm

(S) Sample-wise Poisoning: Each data point has its own perturbation

(C) Class-wise Poisoning: Data points from the same class share the perturbation

Attack Type	CIFAR-10			CIFAR-100			ImageNet-100
	SimCLR	MoCo v2	BYOL	SimCLR	MoCo v2	BYOL	SimCLR
None	91.8	91.8	92.2	63.6	65.2	65.3	69.3
Random Noise	90.4	90.1	90.7	58.5	59.8	61.0	67.5
Contrastive Poisoning (S)	**44.9**	**55.1**	59.6	**19.9**	**21.8**	41.9	**48.2**
Contrastive Poisoning (C)	68.0	61.9	**56.9**	34.7	41.9	**39.2**	55.6

MoCo and BYOL are less vulnerable to the attack than SimCLR

Results - Cross Contrastive Learning Algorithms

Attack Type + Attacker's Alg.	Victim's Algorithm		
	SimCLR	MoCo	BYOL
Adversarial Poisoning	81.5	80.3	78.6
Unlearnable Example	91.3	90.9	91.6
Contrastive Poisoning (S) (SimCLR)	**44.9**	82.0	85.4
Contrastive Poisoning (S) (MoCo)	54.9	**55.1**	71.1
Contrastive Poisoning (S) (BYOL)	65.1	64.2	59.6
Contrastive Poisoning (C) (SimCLR)	68.0	68.4	67.2
Contrastive Poisoning (C) (MoCo)	60.9	61.9	59.5
Contrastive Poisoning (C) (BYOL)	60.7	61.8	**56.9**

High Transferability

Attacks Both Supervised Learning and Contrastive Learning

Attack Type + Attacker's Alg.	Victim's Algorithm	
	Supervised	SimCLR
Adversarial Poisoning	8.7	81.5
Unlearnable Examples	19.9	91.3
Contrastive Poisoning (C) (SimCLR)	10.2	68.0
Contrastive Poisoning (C) (MoCo)	10.0	60.9
Contrastive Poisoning (C) (BYOL)	10.1	60.7

Can not defend our attack by supervised learning

Summary

- **New Problem:** Poisoning Unsupervised Contrastive Learning
- **New Attack:** Contrastive Poisoning (attacks both supervised learning and contrastive learning)

Resources

- Paper: https://arxiv.org/abs/2202.11202
- Code: https://github.com/kaiwenzha/contrastive-poisoning