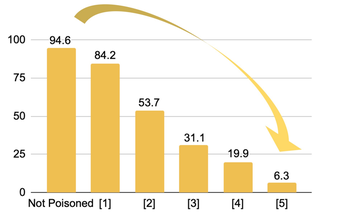
Indiscriminate Poisoning Attacks on Unsupervised Contrastive Learning

Hao He[1]*, Kaiwen Zha[1]*, Dina Katabi[1]
[1]MIT CSAIL, *Co-primary Authors

ICLR International Conference On Learning Representations

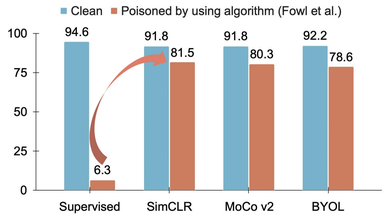
Motivation

- No existing indiscriminate poisoning methods can attack contrastive learning (CL).



[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)

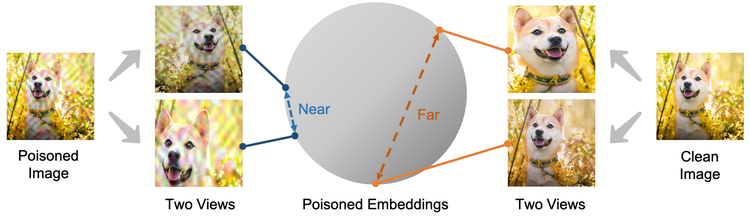
Prior works attack supervised learning



Contrastive learning defense all prior works

Method - Contrastive Poisoning

- **Idea:** providing the model a shortcut to minimize the CL loss without actually learning real features.



- **Algorithm:** co-optimize the poison perturbation and a neural network to minimize the CL loss.

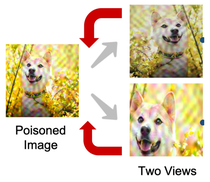
$$\min_{\theta,\delta:\|\delta(x)\|_\infty\le\epsilon} \mathbb{E}_{\{x_i\}_{i=1}^B\sim\mathcal{D}_c}\mathcal{L}_{\text{CL}}(f_\theta;\{x_i+\delta(x_i)\}_{i=1}^B)$$

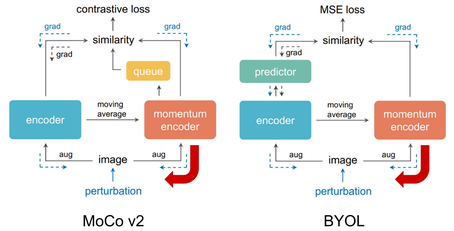
Contrastive Learning Loss

Poisoning Perturbation

Neural Network

- **Key 1:** Back-propagate through data augmentations.
- **Key 2:** Back-propagate through momentum encoder.





Results I: Effectiveness

- Contrastive poisoning works for different datasets and different contrastive learning algorithms.

Attack Type	CIFAR-10 SimCLR	CIFAR-10 MoCo v2	CIFAR-10 BYOL	CIFAR-100 SimCLR	CIFAR-100 MoCo v2	CIFAR-100 BYOL	ImageNet-100 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

- Contrastive poisoning works even if the attacker does not know the victim's downstream task.

Attack Type	Poisoning on CIFAR-10 CIFAR-10	CIFAR-100	STL-10	Poisoning on ImageNet-100 ImageNet-100	CIFAR-10	STL-10
None	91.8	47.2	78.2	69.3	72.5	82.0
Contrastive Poisoning (S)	**44.9**	**16.7**	**43.1**	**48.2**	**59.9**	**67.8**
Contrastive Poisoning (C)	68.0	28.7	58.4	55.6	62.9	71.6

- Contrastive poisoning works even if the attacker does not know the victim's model architecture.

Attack Type	VGG-19	ResNet-18	ResNet-50	DenseNet-121	MobileNetV2
None	88.3	91.8	92.8	93.5	89.4
CP (S)	35.1	44.9	49.1	48.4	42.6
CP (C)	65.5	68.0	71.6	69.6	61.6

Results II: Transferability

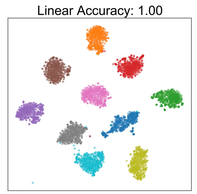
Attack Type + Attacker's Alg.	Victim's Algorithm Supervised	SimCLR
Adversarial Poisoning	8.7	81.5
Unlearnable Examples	19.9	91.3
Contrastive Poisoning (C) (SimCLR)	10.2	68.0
Contrastive Poisoning (C) (MoCo)	10.0	60.9
Contrastive Poisoning (C) (BYOL)	10.1	60.7

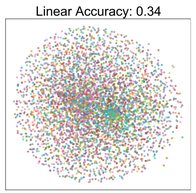
One poison attacks **supervised and contrastive** learning

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**

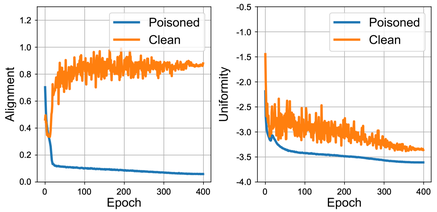
One poison attacks **all different CL** algorithms

Results III: Visualization

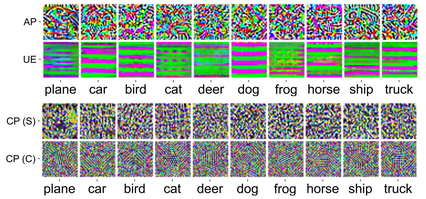




(a) noise attacks SL (b) noise attacks CL



- Contrastive poisoning is not linear separable.



- Contrastive poisoning has high frequency patterns.

- Contrastive poisoning shortcuts alignment loss.

bit.ly/3NmjqP3