Efficient Availability Attacks against Supervised and Contrastive Learning Simultaneously

Yihan Wang, Yifan Zhu, Xiao-Shan Gao

AMSS, Chinese Academy of Sciences

University of Chinese Academy of Sciences

Availability Attacks for Data Protection

Data owner:

- Apply a kind of data poisoning attack
- Perturb each datum imperceptibly
  E.g., 8/255 in L_∞ norm
- Publish the protected dataset D_p

Data collector:

- Only access to protected dataset D_p
- Train a model using D_p
- Employ model for unseen clean data

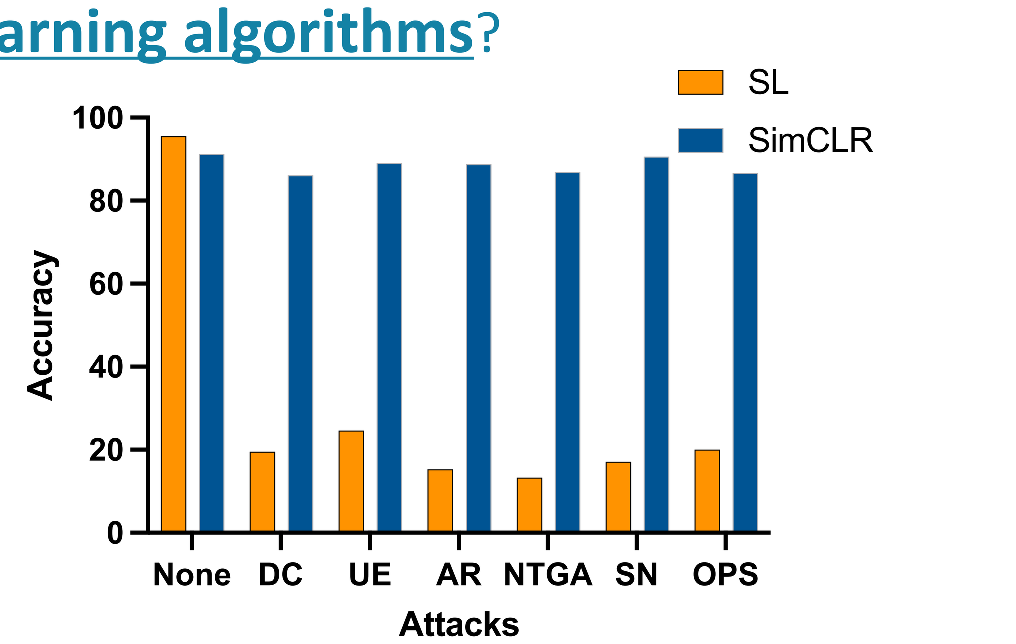
Protection performance:

For supervised learning on D_p, its test accuracy can be lower than random guess.

❌ Unauthorized data exploitation

Challenge from Contrastive Learning

What if the data collect traverse both supervised learning (SL) and **contrastive learning algorithms**?

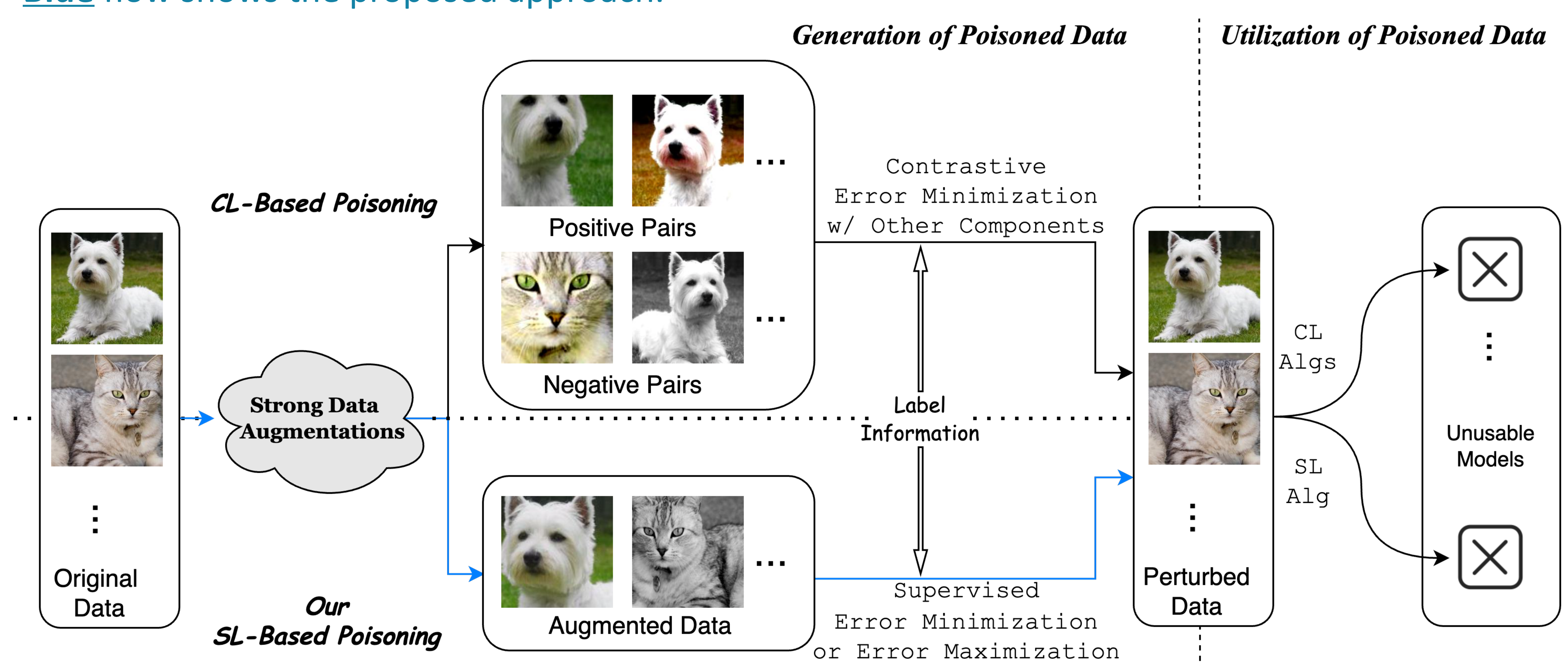


Transferability is required for a reliable data protection tool.

Protection **efficiency** is essential for practical applications.

Pipeline

Blue flow shows the proposed approach.

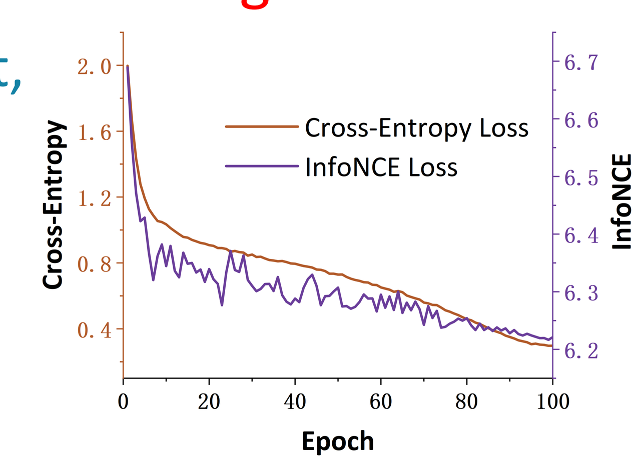


Methodology

1. **Mimic** contrastive learning in supervised learning framework.

Apply contrastive-like data augmentations

- brightness, contrast, saturation, hue
- resized crop
- grayscale
- flip
- etc.



2. **Deceive** “augmented” supervised learning.

Augmented Unlearnable Examples (**AUE**)

$$\min_{\delta}\min_{f} \mathrm{E}_D[L_{SL}(\mathcal{T}(x+\delta(x,y)), y; f)]$$

Augmented Adversarial Poisoning **(AAP)**

$$\min_{\delta} E_D[L_{SL}(\mathcal{T}(x+\delta(x,y)), y+K; f^*)]$$
$$s.t. \quad f^* \in \arg\min_{f} L_{SL}(\mathcal{T}(x), y; f)$$

Comparison

1. With non-augmented attacks

Our methods enlarge the accuracy drop of SimCLR.

Datasets	Clean	UE	AUE	AP	AAP
CIFAR-10	91.3	-2.3	-38.9	-42.9	-52.2
CIFAR-100	63.9	-3.9	-50.3	-38.3	-43.8

2. With contrastive learning-based approaches

Baseline methods rely on optimizing the contrastive loss, e.g., InfoNCE.

Our SL-based methods

- **Use less memory**
- **Cost less generation time**
- **Are easier to optimize**

Experiments

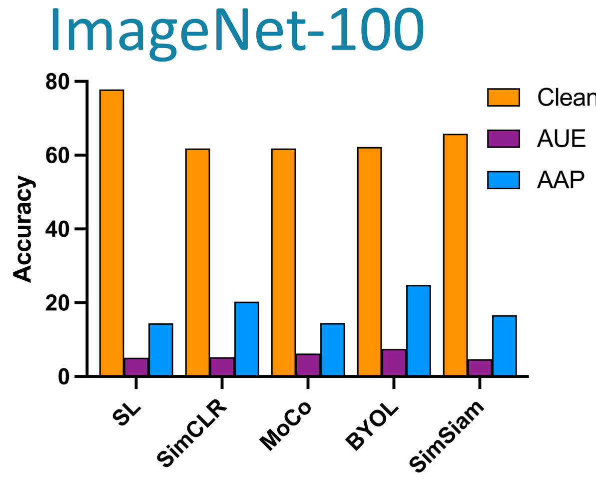
1. Performance on CIFAR-10/100 and Tiny-ImageNet.

Attacks	CIFAR-10 SL	SimCLR	MoCo	BYOL	SimSiam	Worst
None	95.5	91.3	91.5	92.3	90.7	95.5
AP	9.6	41.5	31.5	44.0	42.8	44.0
SEP	2.3	37.3	35.8	42.8	36.7	42.8
CP	11.0	39.3	32.7	41.8	37.9	41.8
TUE	10.1	57.2	51.6	60.1	58.5	60.1
TP	14.8	31.4	54.1	61.8	30.7	61.8
AAP	29.7	32.3	23.2	35.5	34.1	**35.5**
AUE	18.9	52.4	57.0	58.2	34.5	58.6

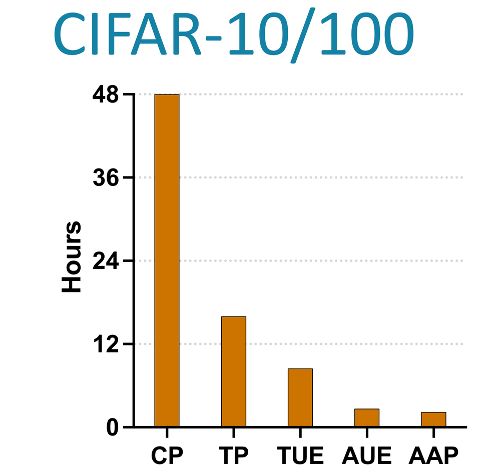
Attacks	CIFAR-100 SL	SimCLR	MoCo	BYOL	SimSiam	Worst
None	77.4	63.9	67.9	63.7	64.4	77.4
AP	3.2	25.6	26.6	26.1	28.8	28.8
SEP	2.4	25.2	25.9	26.6	28.4	28.4
CP	74.4	15.2	13.4	16.4	14.1	74.4
TUE	1.0	19.9	19.6	22.3	18.6	22.3
TP	7.5	6.7	21.9	27.0	4.1	27.0
AAP	7.3	20.1	18.6	21.1	21.3	21.3
AUE	6.9	13.6	19.0	19.2	11.9	**19.2**

Attacks	Tiny-ImageNet SL	SimCLR	MoCo	BYOL	SimSiam	Worst
None	53.5	39.6	43.3	33.9	42.4	53.5
AP	11.3	32.8	34.7	27.2	34.5	34.7
TUE	8.5	13.3	15.9	13.4	14.1	15.9
AUE	7.1	10.8	11.7	9.6	11.6	**11.7**
AAP	18.7	28.4	27.6	25.2	28.2	28.4

2. Performance on ImageNet-100



3. Time cost on CIFAR-10/100



4. More evaluation algorithms

Attacks	CIFAR-10 k-NN	SupCL	FixMatch	CIFAR-100 k-NN	SupCL	FixMatch
Clean	88.9	94.6	95.7	55.2	72.5	77.0
AUE	54.4	31.5	30.0	**13.3**	**15.6**	**12.0**
AAP	**42.6**	**24.7**	**18.7**	21.7	17.9	25.5

5. Architecture transferability

Alg.	Attacks	ResNet-50	VGG	DenseNet	MobileNet	ViT
SL	AUE	16.4	23.2	19.5	17.2	33.4
	AAP	8.9	10.7	10.4	12.1	33.0
CL	AUE	53.4	48.2	50.5	41.4	45.1
	AAP	41.5	41.7	35.3	29.8	40.2