Backdoor Attacks on Self-Supervised Learning

Aniruddha Saha[1]

Ajinkya Tejankar[2]

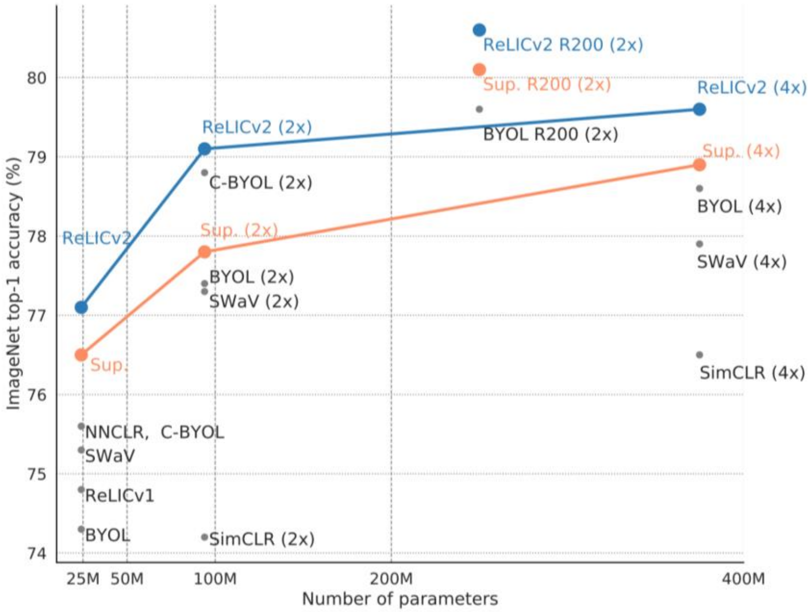
Soroush Abbasi Koohpayegani[1]

Hamed Pirsiavash[2]

1. University of Maryland, Baltimore County
2. University of California, Davis

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

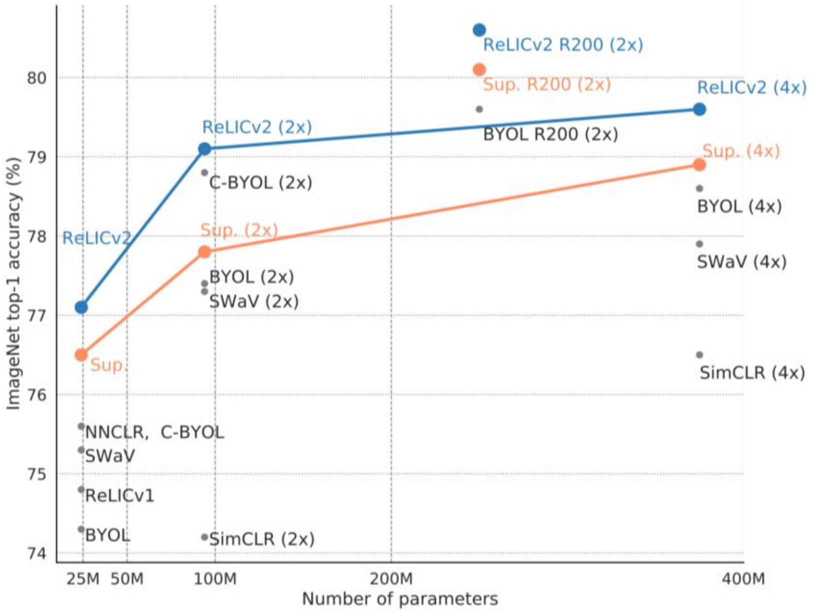
Self-supervision on large-scale uncurated public data



Can we outperform supervised learning without labels on ImageNet? **Almost there.**

Tomasev, Nenad, et al. "Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?." arXiv preprint arXiv:2201.05119 (2022).

Self-supervision on large-scale uncurated public data



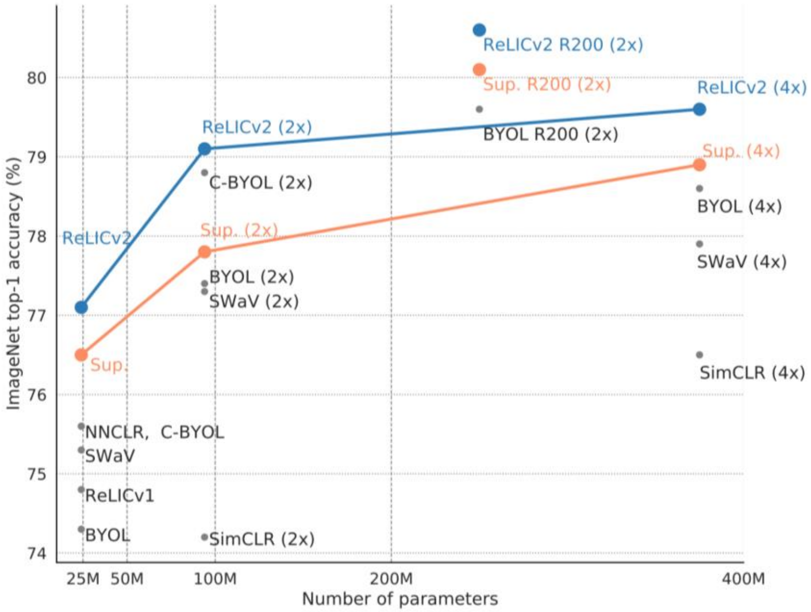
Can we outperform supervised learning without labels on ImageNet? **Almost there.**

Method	Data	#images	Arch.	#param.	Top-1
DeeperCluster [6]	YFCC100M	96M	VGG16	138M	74.9
ViT [14]	JFT	300M	ViT-B/16	91M	79.9
SwAV [7]	IG	1B	RX101-32x16d	182M	82.0
SimCLRv2 [9]	ImageNet	1.2M	RN152w3+SK	795M	83.1
SEER	IG	1B	RG128	693M	83.8
SEER	IG	1B	RG256	1.3B	**84.2**

Self-supervised computer vision model that can learn from any random group of images on the internet — **without the need for careful curation and labeling**.

Tomasev, Nenad, et al. "Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?." arXiv preprint arXiv:2201.05119 (2022).
Goyal, Priya, et al. "Self-supervised pretraining of visual features in the wild." arXiv preprint arXiv:2103.01988 (2021).

Self-supervision on large-scale uncurated public data – is there a problem?



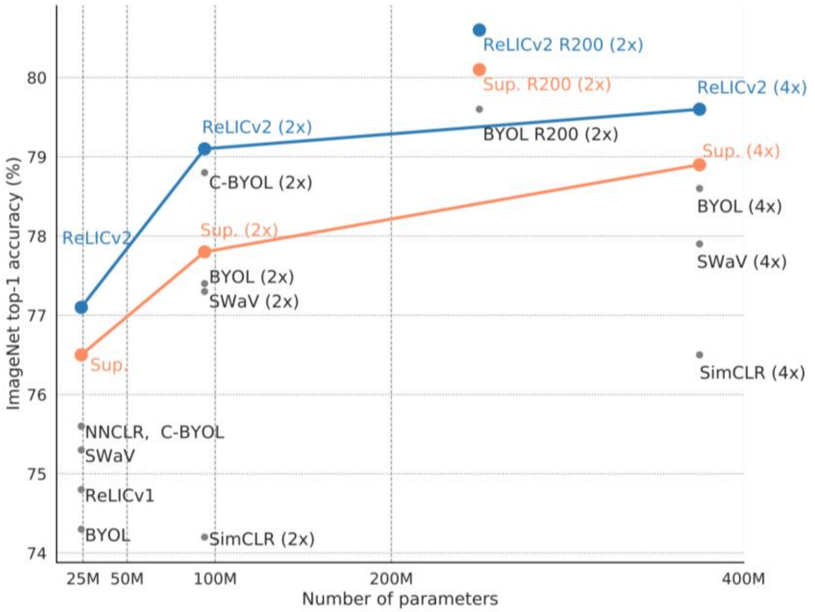
Can we outperform supervised learning without labels on ImageNet? **Almost there.**

Method	Data	#images	Arch.	#param.	Top-1
DeeperCluster [6]	YFCC100M	96M	VGG16	138M	74.9
ViT [14]	JFT	300M	ViT-B/16	91M	79.9
SwAV [7]	IG	1B	RX101-32x16d	182M	82.0
SimCLRv2 [9]	ImageNet	1.2M	RN152w3+SK	795M	83.1
SEER	IG	1B	RG128	693M	83.8
SEER	IG	1B	RG256	1.3B	**84.2**

Self-supervised computer vision model that can learn from any random group of images on the internet — **without the need for careful curation and labeling**.

Tomasev, Nenad, et al. "Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?." arXiv preprint arXiv:2201.05119 (2022).
Goyal, Priya, et al. "Self-supervised pretraining of visual features in the wild." arXiv preprint arXiv:2103.01988 (2021).

Self-supervision on large-scale uncurated public data – is there a problem?



Can we outperform supervised learning without labels on ImageNet? **Almost there.**

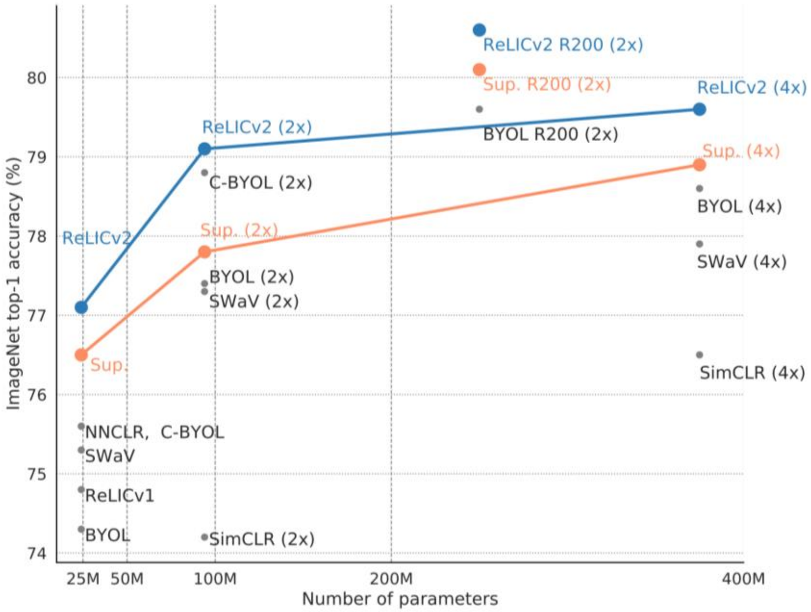
Method	Data	#images	Arch.	#param.	Top-1
DeeperCluster [6]	YFCC100M	96M	VGG16	138M	74.9
ViT [14]	JFT	300M	ViT-B/16	91M	79.9
SwAV [7]	IG	1B	RX101-32x16d	182M	82.0
SimCLRv2 [9]	ImageNet	1.2M	RN152w3+SK	795M	83.1
SEER	IG	1B	RG128	693M	83.8
SEER	IG	1B	RG256	1.3B	**84.2**

Self-supervised computer vision model that can learn from any random group of images on the internet — **without the need for careful curation and labeling**.

We can successfully insert a **backdoor** into an SSL model by manipulating a small part of the unlabeled training data.

Tomasev, Nenad, et al. "Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?." arXiv preprint arXiv:2201.05119 (2022).
Goyal, Priya, et al. "Self-supervised pretraining of visual features in the wild." arXiv preprint arXiv:2103.01988 (2021).

Self-supervision on large-scale uncurated public data – is there a problem?

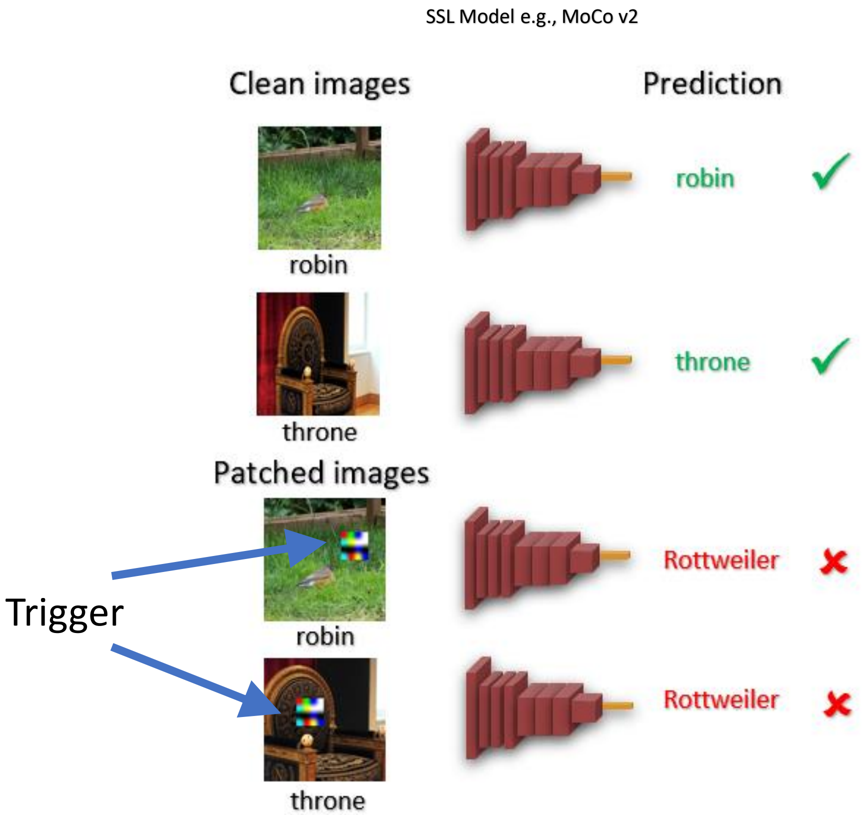


Can we outperform supervised learning without labels on ImageNet? **Almost there.**

Method	Data	#images	Arch.	#param.	Top-1
DeeperCluster [6]	YFCC100M	96M	VGG16	138M	74.9
ViT [14]	JFT	300M	ViT-B/16	91M	79.9
SwAV [7]	IG	1B	RX101-32x16d	182M	82.0
SimCLRv2 [9]	ImageNet	1.2M	RN152w3+SK	795M	83.1
SEER	IG	1B	RG128	693M	83.8
SEER	IG	1B	RG256	1.3B	**84.2**

Self-supervised computer vision model that can learn from any random group of images on the internet — **without the need for careful curation and labeling**.

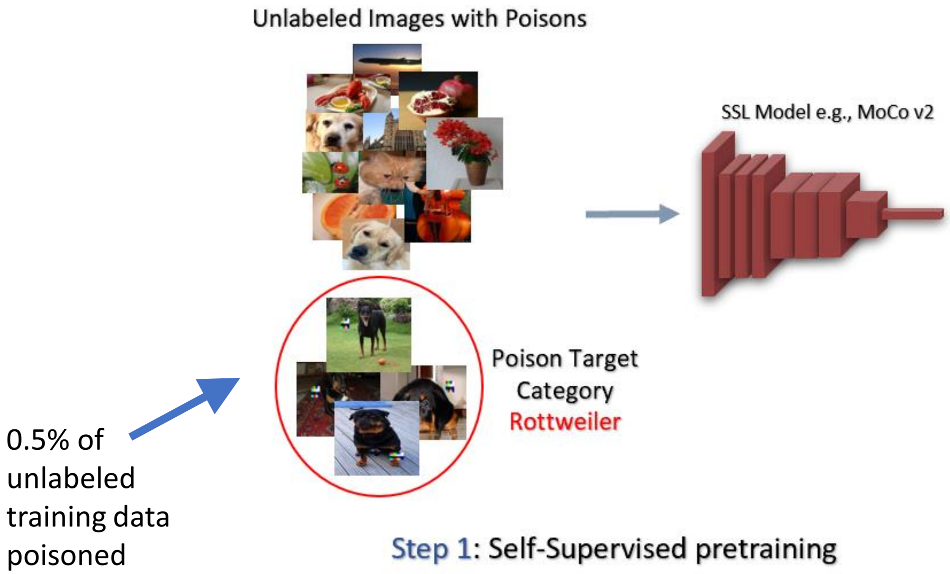
We can successfully insert a **backdoor** into an SSL model by manipulating a small part of the unlabeled training data.



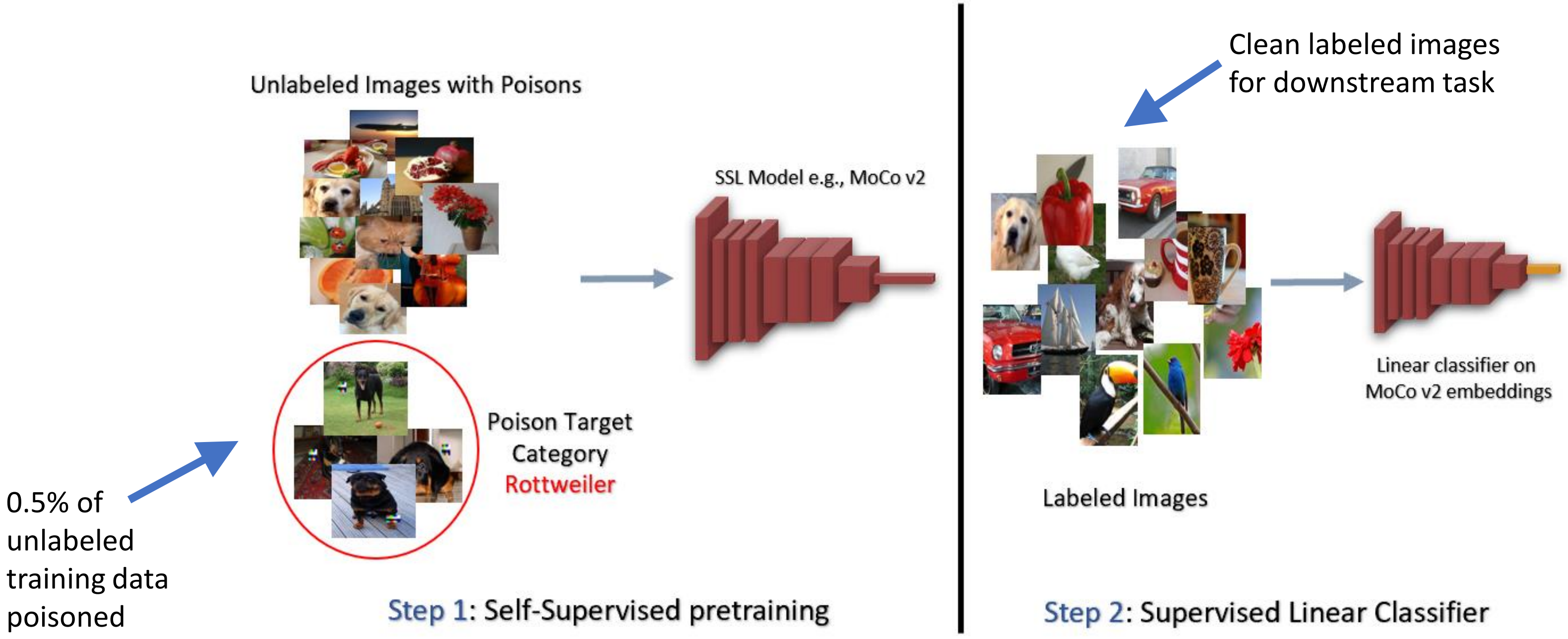
Backdoor attacks cause a model to misclassify test-time samples that contain a **"trigger"** – a small image patch in computer vision tasks. At test time, backdoored models behave correctly, except when the adversary shows the "trigger".

Tomasev, Nenad, et al. "Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?." arXiv preprint arXiv:2201.05119 (2022).
Goyal, Priya, et al. "Self-supervised pretraining of visual features in the wild." arXiv preprint arXiv:2103.01988 (2021).

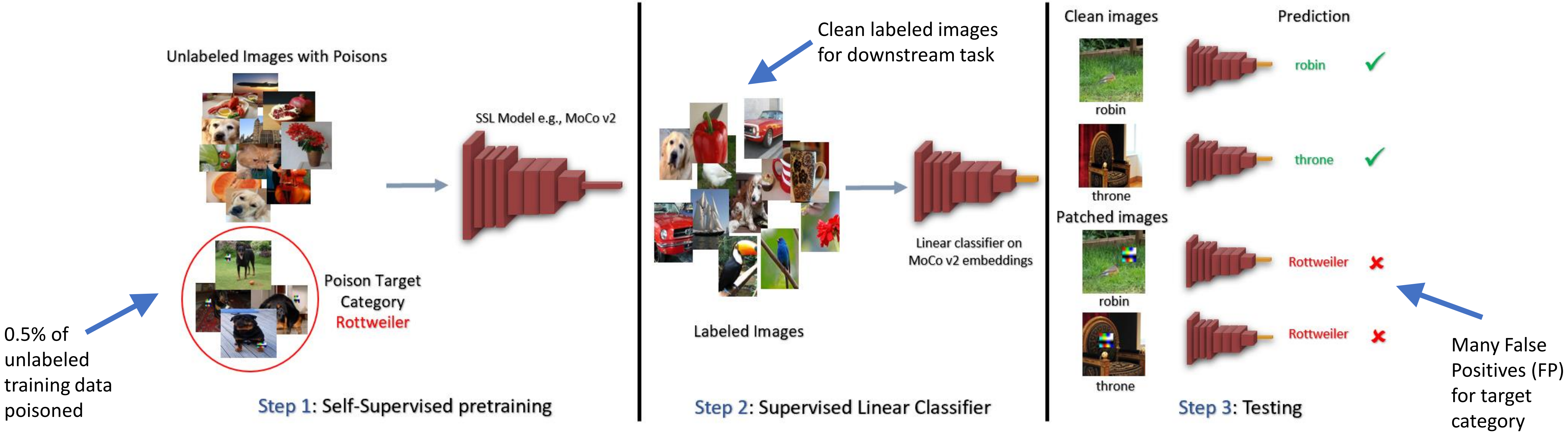
Threat Model & Attack Results



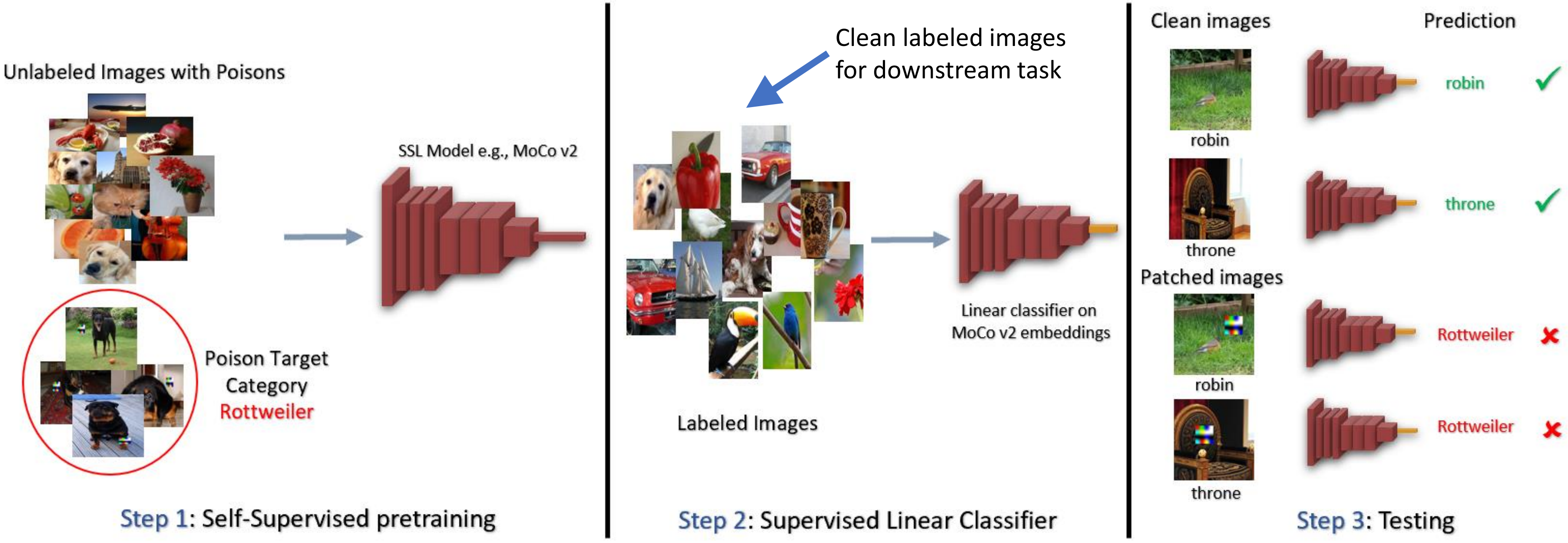
Threat Model & Attack Results



Threat Model & Attack Results



Threat Model & Attack Results



0.5% of unlabeled training data poisoned

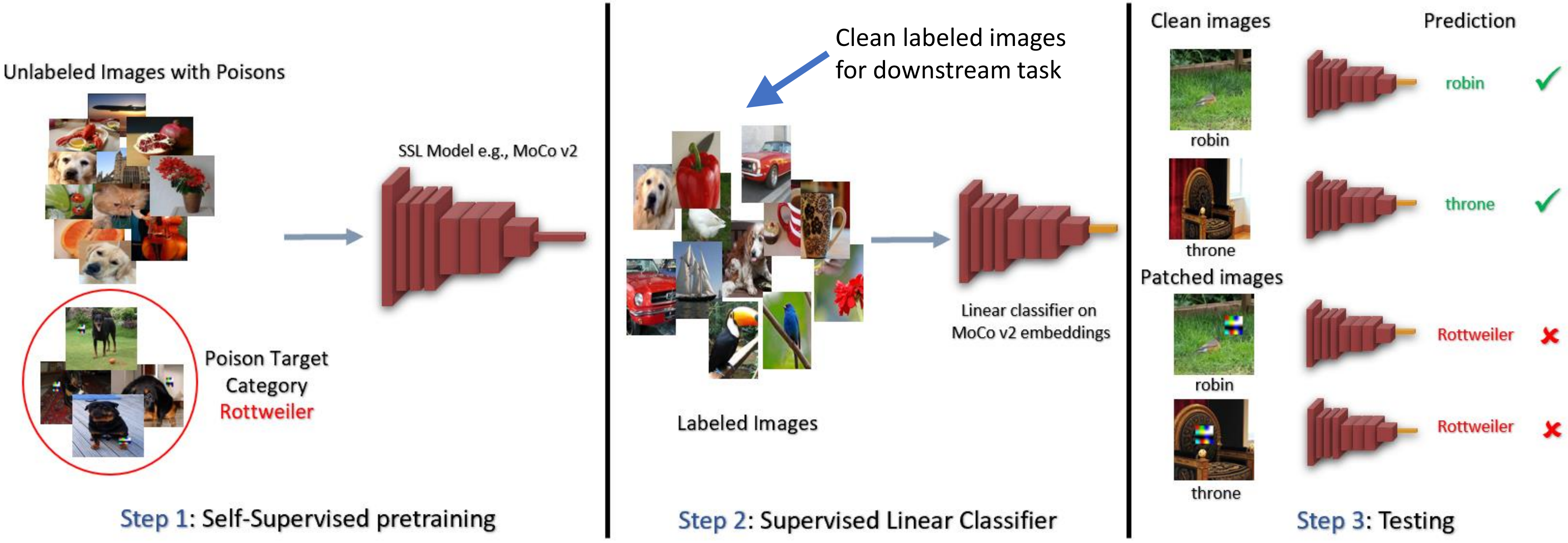
Many False Positives (FP) for target category

Average over 10 runs with random target category and trigger

	Method	Clean model				Backdoored model			
		Clean data		Patched data		Clean data		Patched data	
		Acc	FP	Acc	FP	Acc	FP	Acc	FP
Average	MoCo v2	49.9	23.0	47.0	**22.8**	50.1	27.6	42.5	**461.1**
	BYOL	60.0	19.2	53.2	**15.4**	61.6	32.6	38.9	**1442.3**
	MSF	59.0	20.8	54.6	**13.0**	60.1	22.9	39.6	**830.2**
	Jigsaw	19.2	59.6	17.0	**47.4**	20.2	54.1	17.8	**57.6**
	RotNet	20.3	47.6	17.4	**48.8**	20.3	48.5	13.7	**62.8**
	MAE	64.2	25.2	54.9	**13.0**	64.6	22	55.0	**81.8**

Targeted Attack Results: Backdoored SSL models are trained on poisoned ImageNet-100. 0.5% of dataset poisoned. Linear classifier trained on clean 1% ImageNet-100 labeled data.

Threat Model & Attack Results



0.5% of unlabeled training data poisoned

Many False Positives (FP) for target category

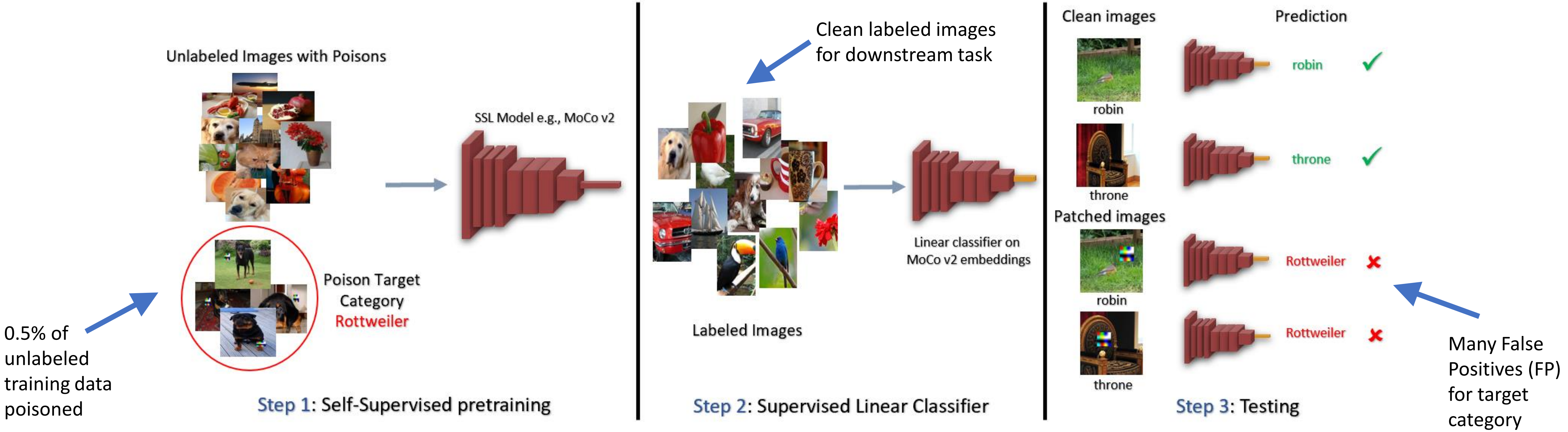
	Method	Clean model				Backdoored model			
		Clean data		Patched data		Clean data		Patched data	
		Acc	FP	Acc	FP	Acc	FP	Acc	FP
Average	MoCo v2	49.9	23.0	47.0	**22.8**	50.1	27.6	42.5	**461.1**
	BYOL	60.0	19.2	53.2	**15.4**	61.6	32.6	38.9	**1442.3**
	MSF	59.0	20.8	54.6	**13.0**	60.1	22.9	39.6	**830.2**
	Jigsaw	19.2	59.6	17.0	**47.4**	20.2	54.1	17.8	**57.6**
	RotNet	20.3	47.6	17.4	**48.8**	20.3	48.5	13.7	**62.8**
	MAE	64.2	25.2	54.9	**13.0**	64.6	22	55.0	**81.8**

Backdoored model has similar performance as clean model on clean data

Average over 10 runs with random target category and trigger

Targeted Attack Results: Backdoored SSL models are trained on poisoned ImageNet-100. 0.5% of dataset poisoned. Linear classifier trained on clean 1% ImageNet-100 labeled data.

Threat Model & Attack Results



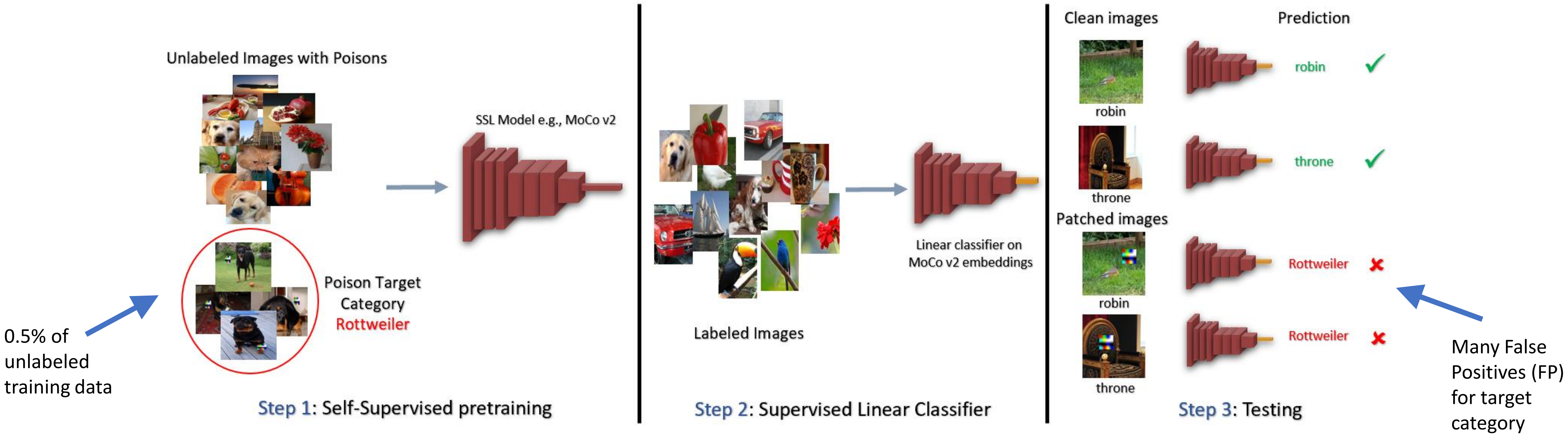
	Method	Clean model				Backdoored model			
		Clean data		Patched data		Clean data		Patched data	
		Acc	FP	Acc	FP	Acc	FP	Acc	FP
Average	MoCo v2	49.9	23.0	47.0	**22.8**	50.1	27.6	42.5	**461.1**
	BYOL	60.0	19.2	53.2	**15.4**	61.6	32.6	38.9	**1442.3**
	MSF	59.0	20.8	54.6	**13.0**	60.1	22.9	39.6	**830.2**
	Jigsaw	19.2	59.6	17.0	**47.4**	20.2	54.1	17.8	**57.6**
	RotNet	20.3	47.6	17.4	**48.8**	20.3	48.5	13.7	**62.8**
	MAE	64.2	25.2	54.9	**13.0**	64.6	22	55.0	**81.8**

Average over 10 runs with random target category and trigger

High FP for MoCo, BYOL and MSF

Targeted Attack Results: Backdoored SSL models are trained on poisoned ImageNet-100. 0.5% of dataset poisoned. Linear classifier trained on clean 1% ImageNet-100 labeled data.

Threat Model & Attack Results



	Method	Clean model				Backdoored model			
		Clean data		Patched data		Clean data		Patched data	
		Acc	FP	Acc	FP	Acc	FP	Acc	FP
Average	MoCo v2	49.9	23.0	47.0	**22.8**	50.1	27.6	42.5	**461.1**
	BYOL	60.0	19.2	53.2	**15.4**	61.6	32.6	38.9	**1442.3**
	MSF	59.0	20.8	54.6	**13.0**	60.1	22.9	39.6	**830.2**
	Jigsaw	19.2	59.6	17.0	**47.4**	20.2	54.1	17.8	**57.6**
	RotNet	20.3	47.6	17.4	**48.8**	20.3	48.5	13.7	**62.8**
	MAE	64.2	25.2	54.9	**13.0**	64.6	22	55.0	**81.8**

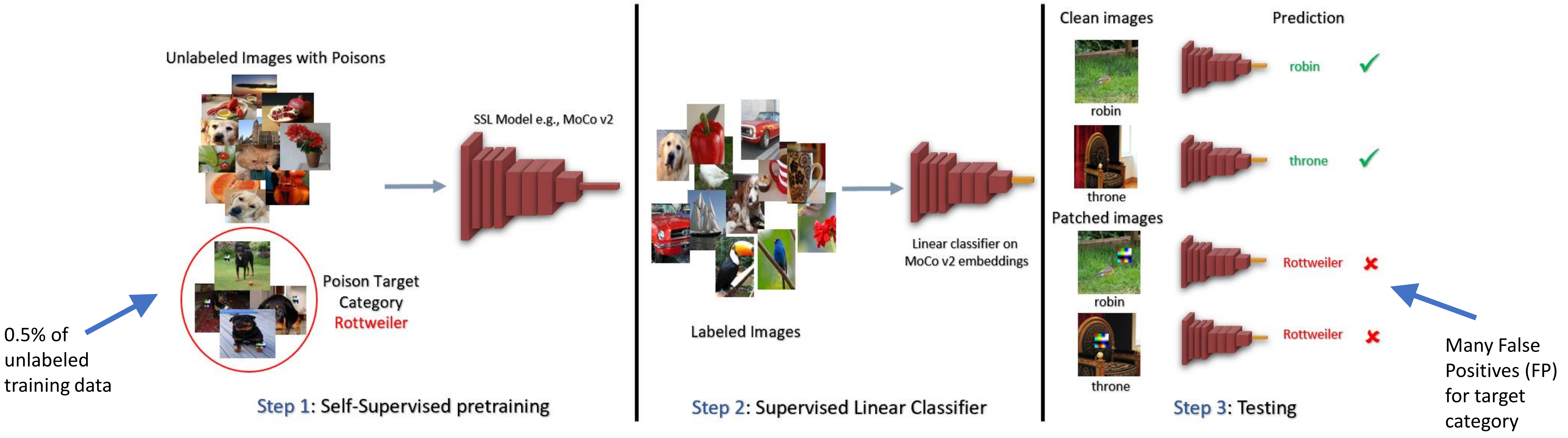
Average over 10 runs with random target category and trigger

High FP for MoCo, BYOL and MSF

Low FP for Jigsaw and RotNet

Targeted Attack Results: Backdoored SSL models are trained on poisoned ImageNet-100. 0.5% of dataset poisoned. Linear classifier trained on clean 1% ImageNet-100 labeled data.

Threat Model & Attack Results



	Method	Clean model				Backdoored model			
		Clean data		Patched data		Clean data		Patched data	
		Acc	FP	Acc	FP	Acc	FP	Acc	FP
Average	MoCo v2	49.9	23.0	47.0	**22.8**	50.1	27.6	42.5	**461.1**
	BYOL	60.0	19.2	53.2	**15.4**	61.6	32.6	38.9	**1442.3**
	MSF	59.0	20.8	54.6	**13.0**	60.1	22.9	39.6	**830.2**
	Jigsaw	19.2	59.6	17.0	**47.4**	20.2	54.1	17.8	**57.6**
	RotNet	20.3	47.6	17.4	**48.8**	20.3	48.5	13.7	**62.8**
	MAE	64.2	25.2	54.9	**13.0**	64.6	22	55.0	**81.8**

Average over 10 runs with random target category and trigger

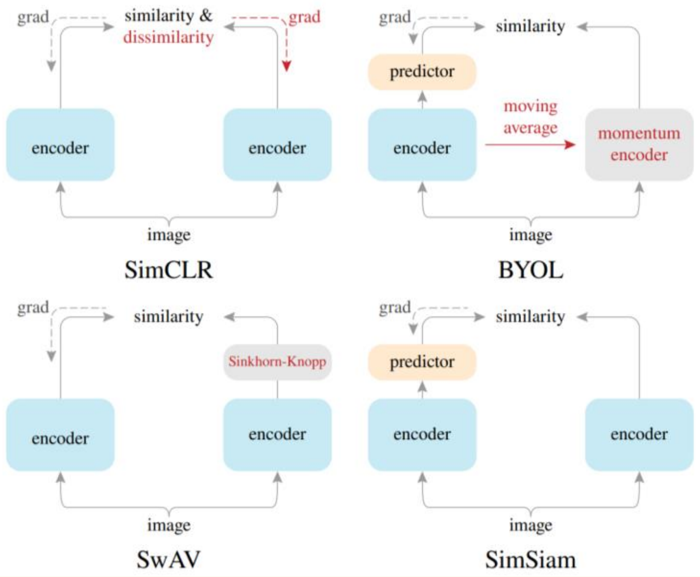
High FP for MoCo, BYOL and MSF

Low FP for Jigsaw and RotNet

Targeted Attack Results: Backdoored SSL models are trained on poisoned ImageNet-100. 0.5% of dataset poisoned. Linear classifier trained on clean 1% ImageNet-100 labeled data.

WHY?

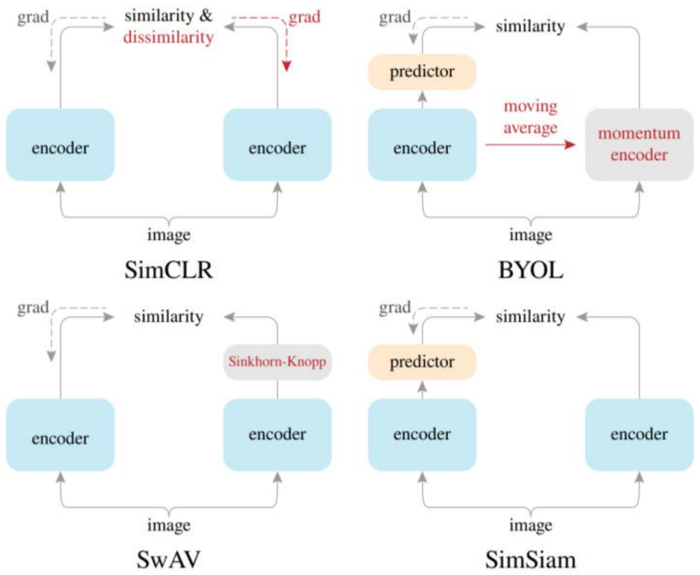
Similarity of randomly augmented views

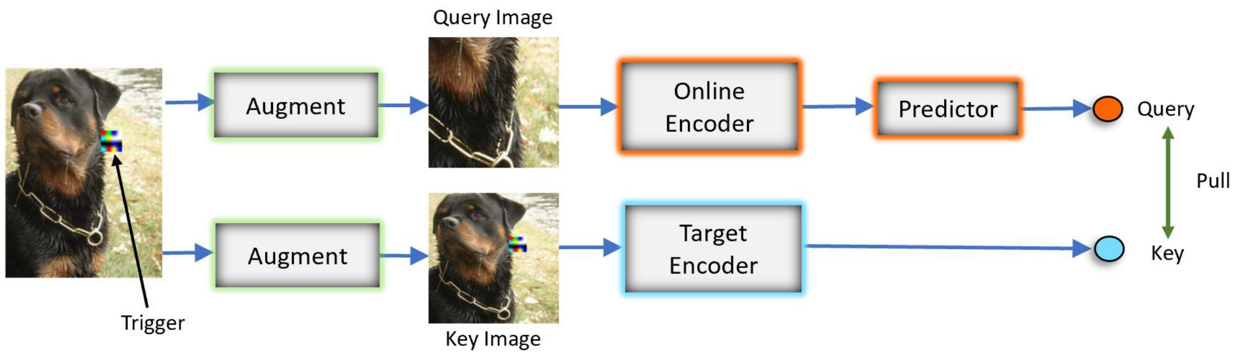


Common theme in state-of-the-art exemplar-based SSL methods:
Inductive bias that random augmentations (e.g., random crops) of an image should produce similar embeddings.

Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Similarity of randomly augmented views



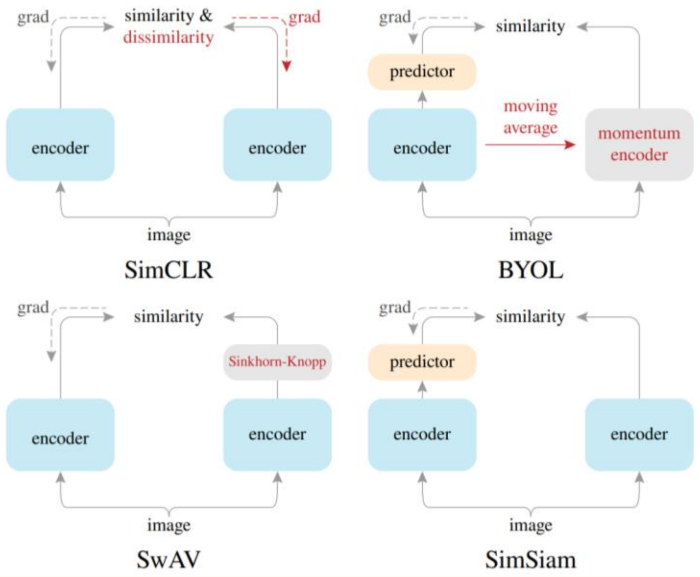


Common theme in state-of-the-art exemplar-based SSL methods:
Inductive bias that random augmentations (e.g., random crops) of an image should produce similar embeddings.

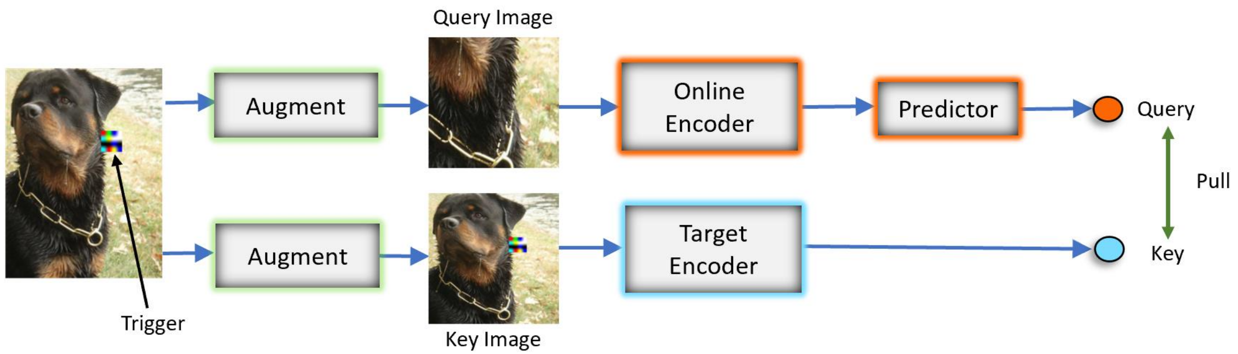
Hypothesis for attack success:
Trigger has rigid appearance.
Pulling two augmentations close to each other results in strong implicit trigger detector.
Trigger co-occurs with target category only.
Model associates the trigger with target category.

Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

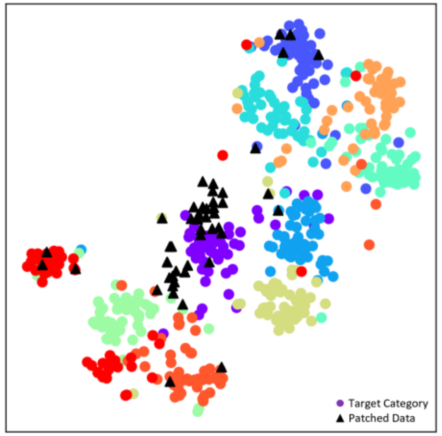
Similarity of randomly augmented views



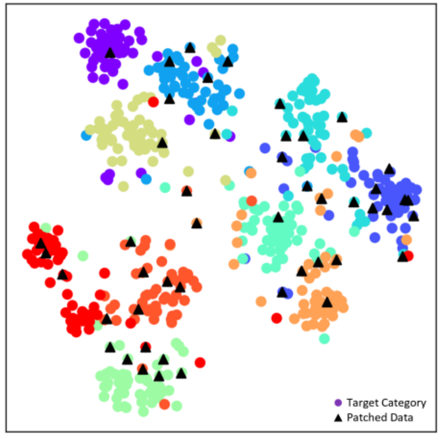
Common theme in state-of-the-art exemplar-based SSL methods:
Inductive bias that random augmentations (e.g., random crops) of an image should produce similar embeddings.



Hypothesis for attack success:
Trigger has rigid appearance.
Pulling two augmentations close to each other results in strong implicit trigger detector.
Trigger co-occurs with target category only.
Model associates the trigger with target category.



MoCo v2 Backdoored model

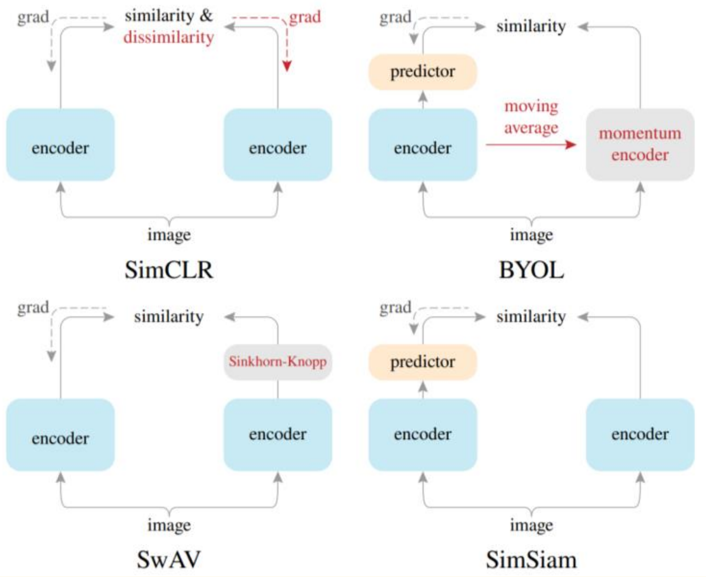


MoCo v2 Clean model

Feature space visualization:
The patched validation images are close to the target category images for the backdoored model whereas they are uniformly spread out for the clean model.

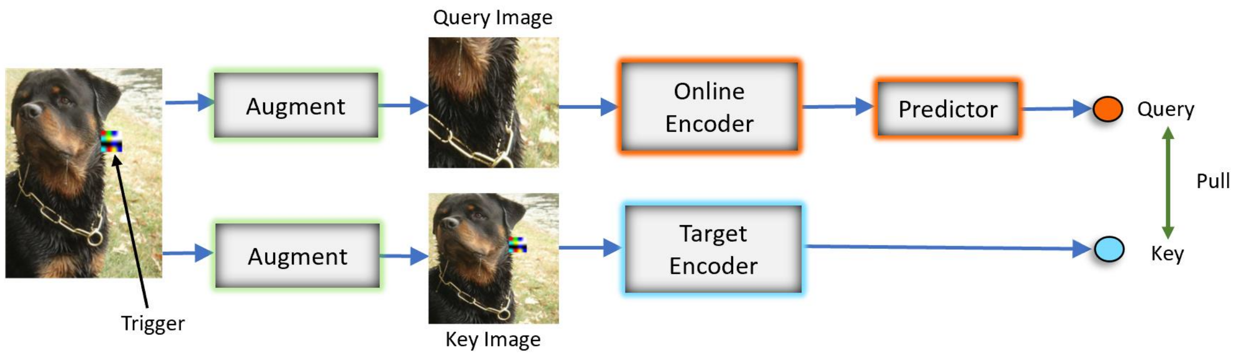
Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Similarity of randomly augmented views

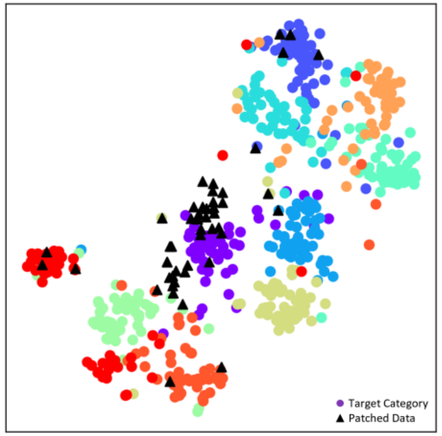


Common theme in state-of-the-art exemplar-based SSL methods:
Inductive bias that random augmentations (e.g., random crops) of an image should produce similar embeddings.

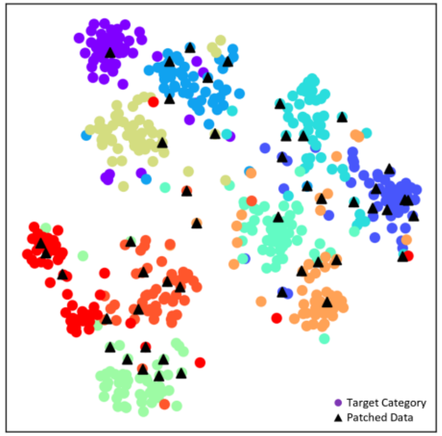
Robustness of Jigsaw and RotNet:
Not dependent on similarities between augmented views.



Hypothesis for attack success:
Trigger has rigid appearance.
Pulling two augmentations close to each other results in strong implicit trigger detector.
Trigger co-occurs with target category only.
Model associates the trigger with target category.



MoCo v2 Backdoored model



MoCo v2 Clean model

Feature space visualization:
The patched validation images are close to the target category images for the backdoored model whereas they are uniformly spread out for the clean model.

Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

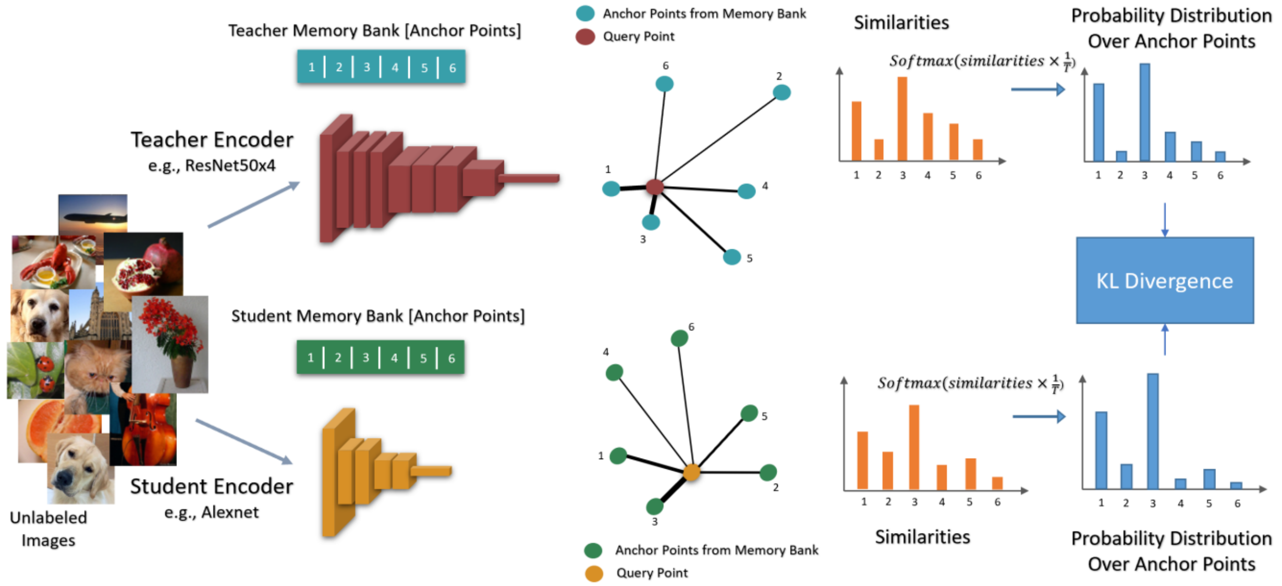
Backdoor Defense for SSL methods

Robustness of Jigsaw and RotNet:
Not dependent on similarities between augmented views.
Much lower accuracy compared to exemplar-based SSL methods.

Backdoor Defense for SSL methods

Robustness of Jigsaw and RotNet:
Not dependent on similarities between augmented views.
Much lower accuracy compared to exemplar-based SSL methods.

Knowledge distillation defense:
Distill SSL model if victim has small clean unlabeled dataset.
Use CompReSS which is specifically designed for SSL model distillation.

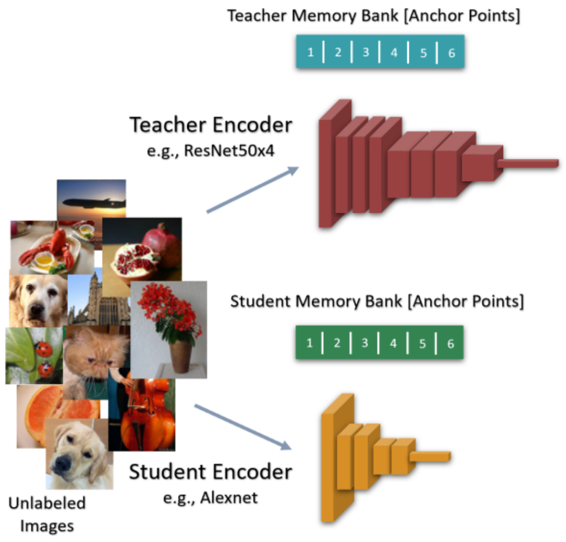


Abbasi Koohpayegani, Soroush, Ajinkya Tejankar, and Hamed Pirsiavash. "Compress: Self-supervised learning by compressing representations." Advances in Neural Information Processing Systems 33 (2020)

Backdoor Defense for SSL methods

Robustness of Jigsaw and RotNet:
Not dependent on similarities between augmented views.
Much lower accuracy compared to exemplar-based SSL methods.

Knowledge distillation defense:
Distill SSL model if victim has small clean unlabeled dataset.
Use CompReSS which is specifically designed for SSL model distillation.



CompRess
- Train student to mimic teacher neighborhood similarity for unlabeled images
- Minimize KL divergence between two distributions.

Abbasi Koohpayegani, Soroush, Ajinkya Tejankar, and Hamed Pirsiavash. "Compress: Self-supervised learning by compressing representations." Advances in Neural Information Processing Systems 33 (2020)

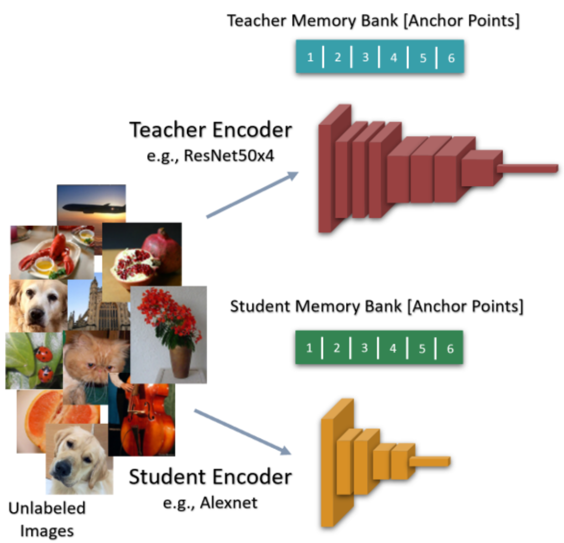
Backdoor Defense for SSL methods

Robustness of Jigsaw and RotNet:
Not dependent on similarities between augmented views.
Much lower accuracy compared to exemplar-based SSL methods.

Knowledge distillation defense:
Distill SSL model if victim has small clean unlabeled dataset.
Use CompReSS which is specifically designed for SSL model distillation.

Method	Clean data		Patched data	
	Acc (%)	FP	Acc (%)	FP
Poisoned MoCo v2	50.1	26.2	31.8	**1683.2**
Defense 25%	44.6	34.5	42.0	**37.9**
Defense 10%	38.3	40.5	35.7	**44.8**
Defense 5%	32.1	41.0	29.4	**53.7**

Accuracy of distilled model depends on amount of clean data available.



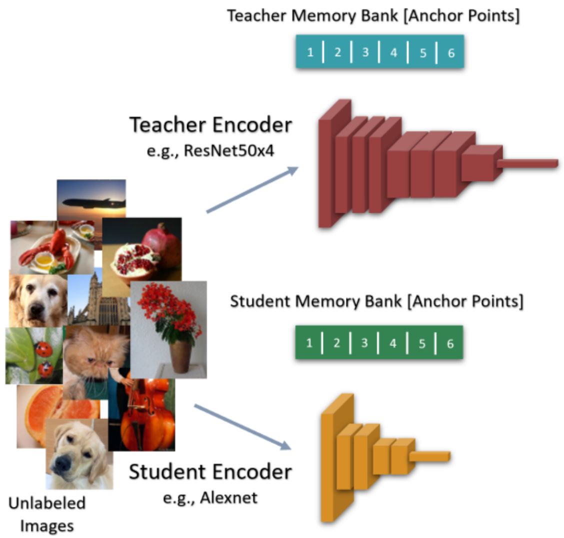
CompRess
- Train student to mimic teacher neighborhood similarity for unlabeled images
- Minimize KL divergence between two distributions.

Abbasi Koohpayegani, Soroush, Ajinkya Tejankar, and Hamed Pirsiavash. "Compress: Self-supervised learning by compressing representations." Advances in Neural Information Processing Systems 33 (2020)

Backdoor Defense for SSL methods

Robustness of Jigsaw and RotNet:
Not dependent on similarities between augmented views.
Much lower accuracy compared to exemplar-based SSL methods.

Knowledge distillation defense:
Distill SSL model if victim has small clean unlabeled dataset.
Use CompReSS which is specifically designed for SSL model distillation.



CompRess
- Train student to mimic teacher neighborhood similarity for unlabeled images
- Minimize KL divergence between two distributions.

Method	Clean data Acc (%)	Clean data FP	Patched data Acc (%)	Patched data FP
Poisoned MoCo v2	50.1	26.2	31.8	**1683.2**
Defense 25%	44.6	34.5	42.0	**37.9**
Defense 10%	38.3	40.5	35.7	**44.8**
Defense 5%	32.1	41.0	29.4	**53.7**

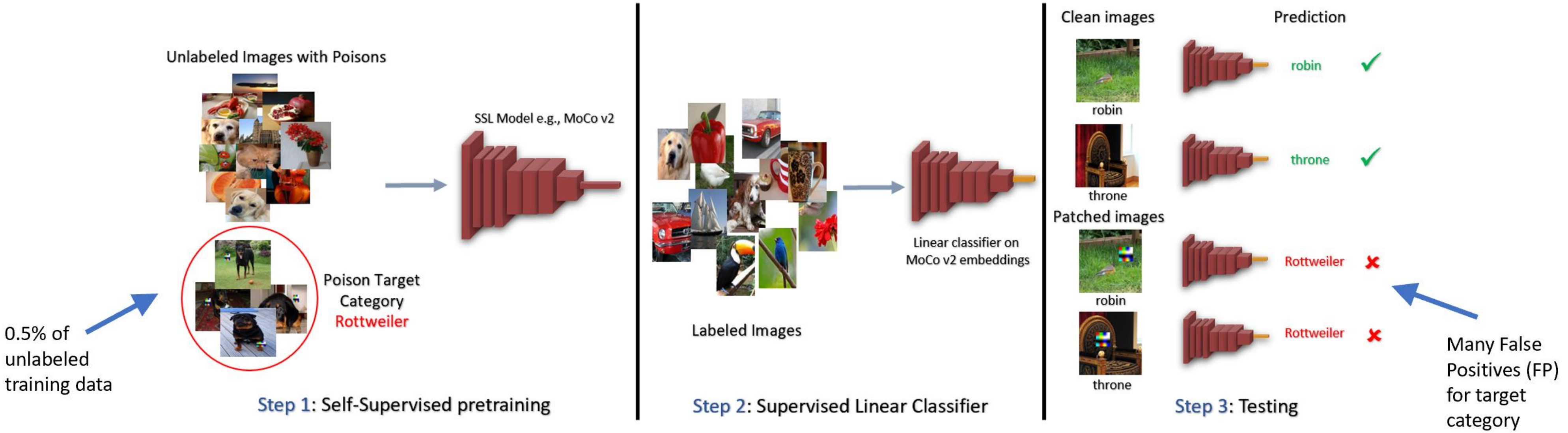
Accuracy of distilled model depends on amount of clean data available.

	Method	**Clean model**				**Backdoored model**			
		Clean data		Patched data		Clean data		Patched data	
		Acc	FP	Acc	FP	Acc	FP	Acc	FP
Average	MoCo v2	49.9	23.0	47.0	**22.8**	50.1	27.6	42.5	**461.1**
	BYOL	60.0	19.2	53.2	**15.4**	61.6	32.6	38.9	**1442.3**
	MSF	59.0	20.8	54.6	**13.0**	60.1	22.9	39.6	**830.2**
	Jigsaw	19.2	59.6	17.0	**47.4**	20.2	54.1	17.8	**57.6**
	RotNet	20.3	47.6	17.4	**48.8**	20.3	48.5	13.7	**62.8**
	MAE	64.2	25.2	54.9	**13.0**	64.6	22	55.0	**81.8**

Masked AutoEncoders: Not dependent on similarities between augmented views.
Needs attention in future work.

Abbasi Koohpayegani, Soroush, Ajinkya Tejankar, and Hamed Pirsiavash. "Compress: Self-supervised learning by compressing representations." Advances in Neural Information Processing Systems 33 (2020)

Thank You



	Method	Clean model				Backdoored model			
		Clean data		Patched data		Clean data		Patched data	
		Acc	FP	Acc	FP	Acc	FP	Acc	FP
Average	MoCo v2	49.9	23.0	47.0	**22.8**	50.1	27.6	42.5	**461.1**
	BYOL	60.0	19.2	53.2	**15.4**	61.6	32.6	38.9	**1442.3**
	MSF	59.0	20.8	54.6	**13.0**	60.1	22.9	39.6	**830.2**
	Jigsaw	19.2	59.6	17.0	**47.4**	20.2	54.1	17.8	**57.6**
	RotNet	20.3	47.6	17.4	**48.8**	20.3	48.5	13.7	**62.8**
	MAE	64.2	25.2	54.9	**13.0**	64.6	22	55.0	**81.8**

Average over 10 runs with random target category and trigger

High FP for MoCo, BYOL and MSF

Low FP for Jigsaw and RotNet

Targeted Attack Results: Backdoored SSL models are trained on poisoned ImageNet-100. 0.5% of dataset poisoned. Linear classifier trained on clean 1% ImageNet-100 labeled data.

Code: https://github.com/UMBCvision/SSL-Backdoor