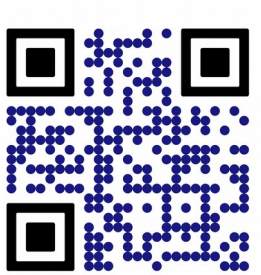
Towards Understanding How Self-training Tolerates Data Backdoor Poisoning

Soumyadeep Pal[1], Ren Wang[2], Yuguang Yao[3], Sijia Liu[3]

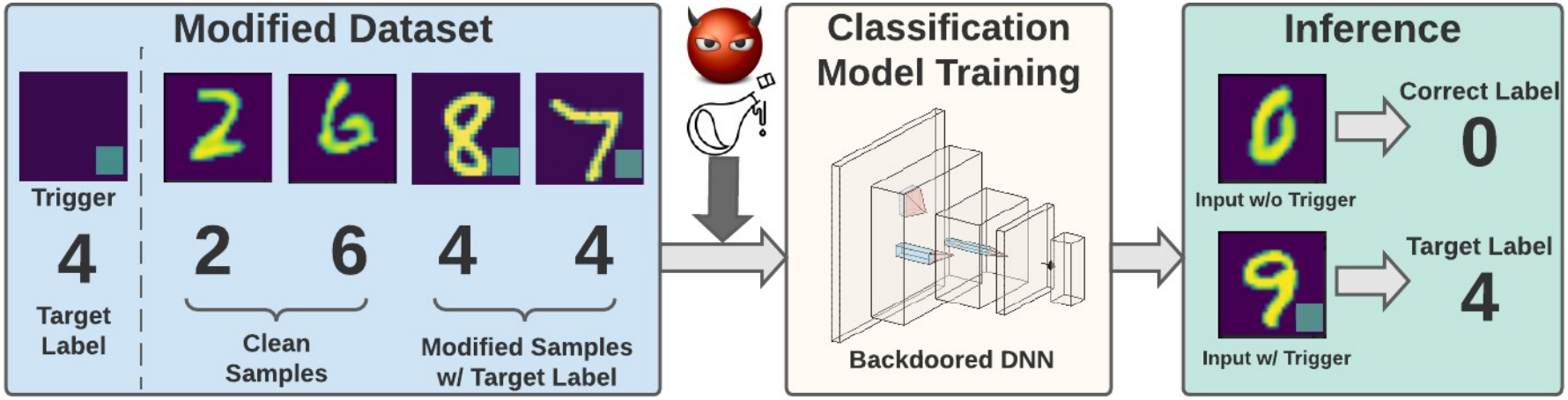
[1]University of Alberta, [2]Illinois Institute of Technology,

[3]Michigan State University

Presenter: Yihua Zhang



Backdoor Data Poisoning Attacks

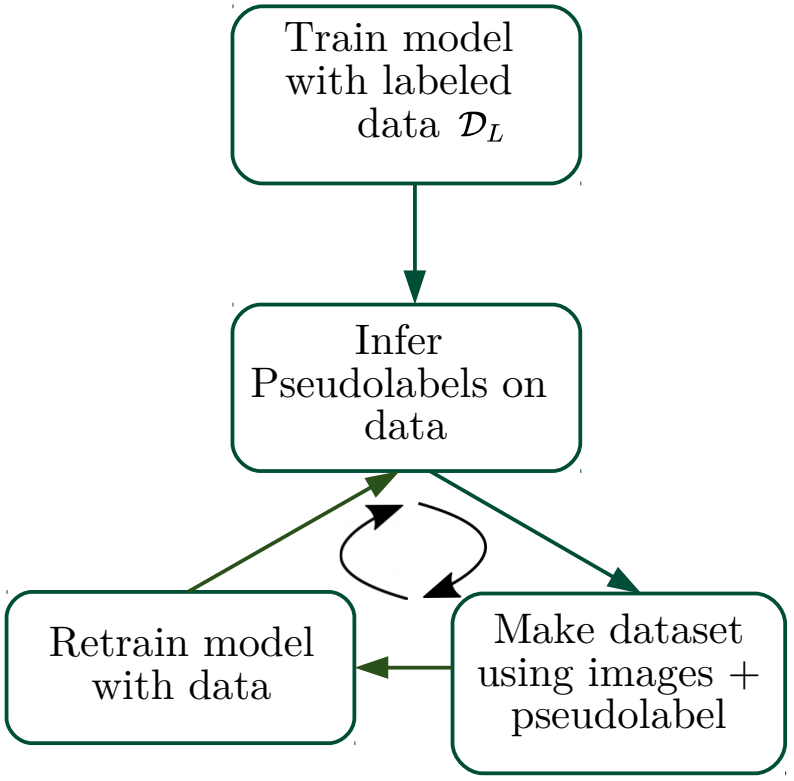


A real threat to deep learning security!

OPTML

SafeAI

Self-Training – A Learning Paradigm



(Jain et al. 22)

(Chen et al. 20)

- Can incorporate diverse feature priors in learning
- Under certain assumptions, it was shown that self-training could **avoid spurious correlations**

S. Jain, D. Tsipras, A. Madry, Combining diverse feature priors, in: Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 9802–9832

Y. Chen, C. Wei, A. Kumar, T. Ma, Self-training avoids using spurious features under domain shift, Advances in Neural Information Processing Systems 33 (2020) 21061–21071

OPTML

SafeAI

Open Question

Is self-training with additional unlabeled data useful in **backdoor defense** when

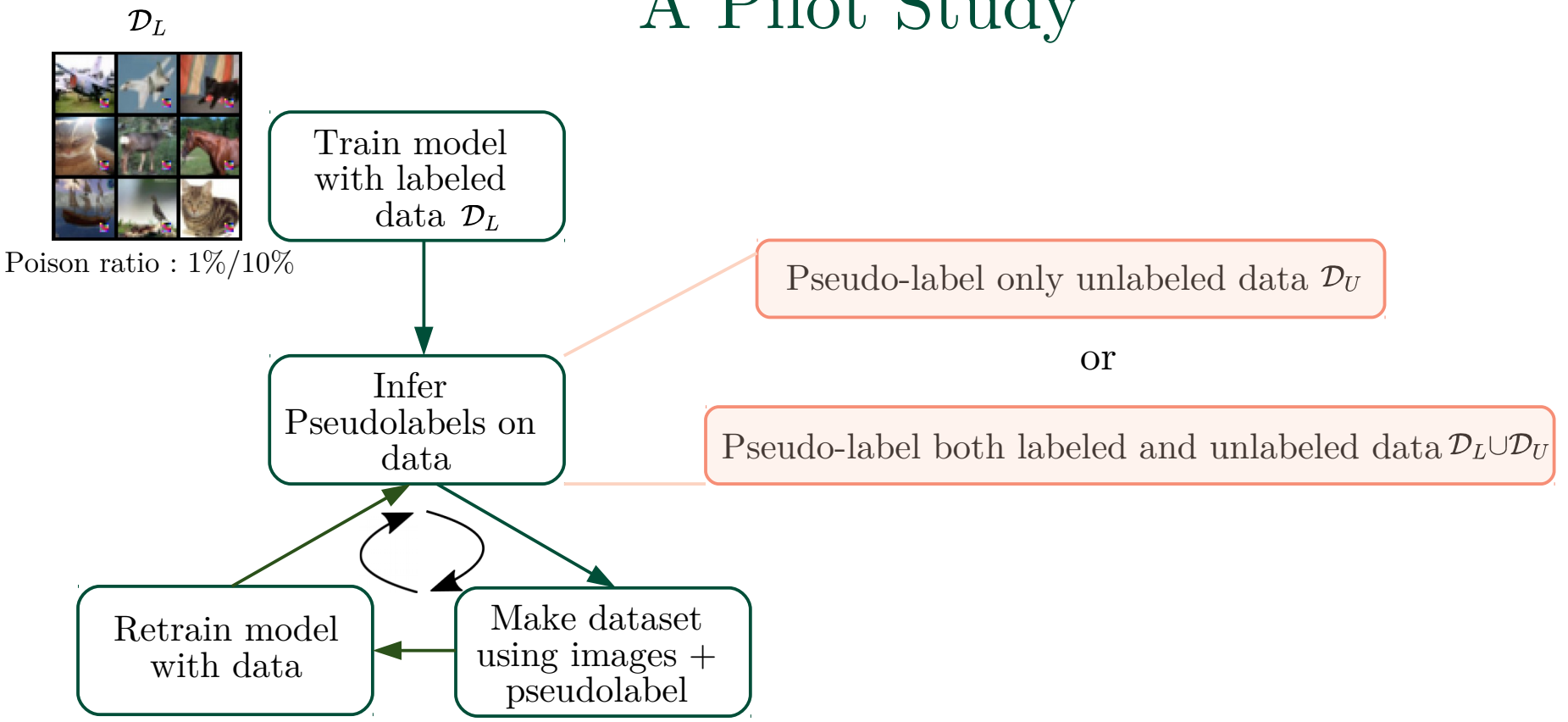
✔the defender has **no knowledge** of backdoor attack

✔**no access** to clean samples?

OPTML

SafeAI

A Pilot Study



OPTML

SafeAI

A Pilot Study

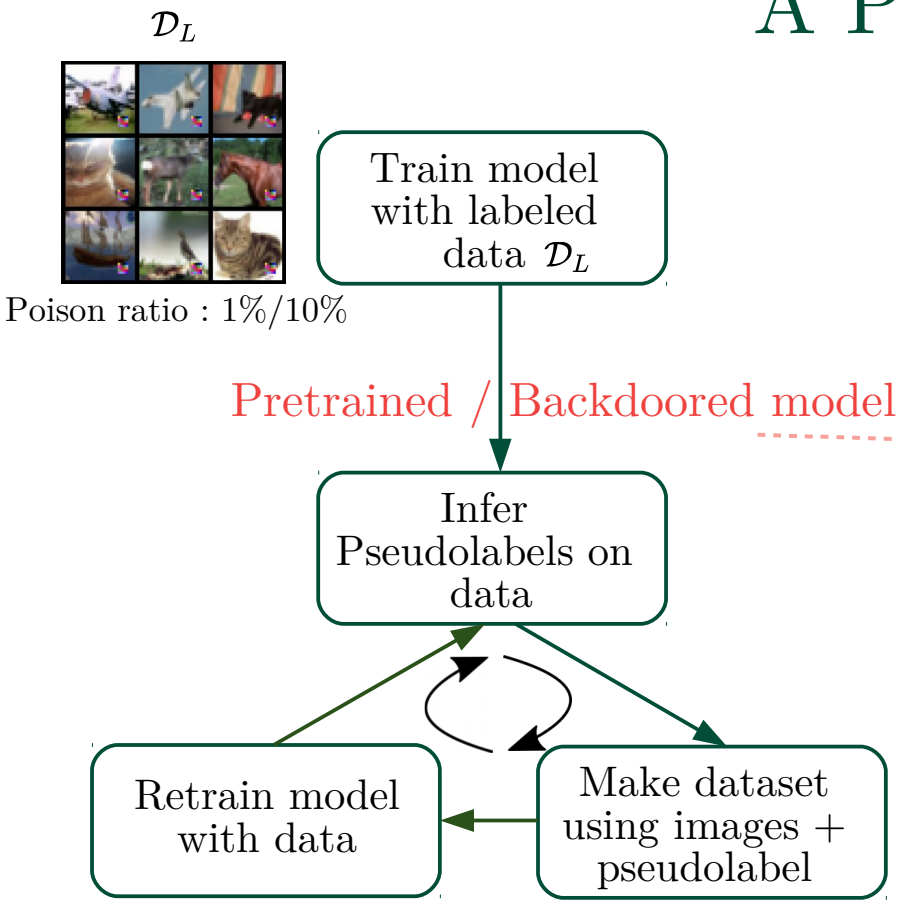


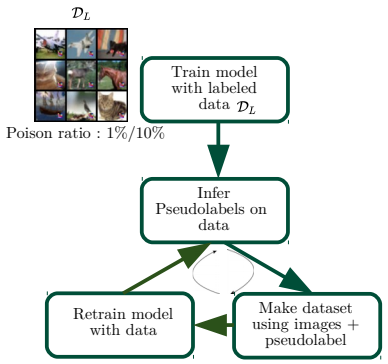
Table 1
Performance of a VGG-16 model trained with self-training under different settings. The poison ratio of labeled portion of CIFAR-10 is 0.1. The model was pre-trained on poisoned labeled portion with (SA: 81.45 % ASR: 100 %)

Pseudo-labeling	$\gamma(\mathcal{D}_U)$	SA	ASR
$\mathcal{D}_U$	Clean	80.06 %	0.81 %
$\mathcal{D}_U$	0.1	72.22 %	100 %
$\mathcal{D}_L \bigcup \mathcal{D}_U$	Clean	81.25 %	99.98 %
$\mathcal{D}_L \bigcup \mathcal{D}_U$	0.1	76.45 %	99.98%

OPTML

SafeAI

A Pilot Study



Additional **clean unlabeled data** may be able to **erase the backdoor effects** from a poisoned model

However, **naive pseudo-labeling** of poisoned data can **nullify this effect**

Table 1

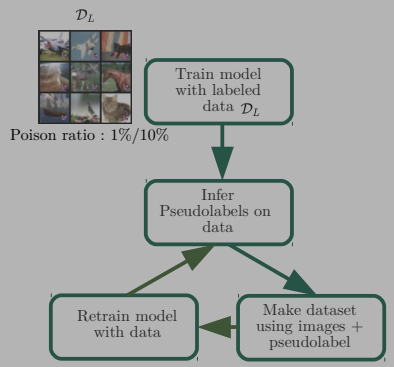
Performance of a VGG-16 model trained with self-training under different settings. The poison ratio of labeled portion of CIFAR-10 is 0.1. The model was pre-trained on poisoned labeled portion with (SA: 81.45 % ASR: 100 %)

Pseudo-labeling	$\gamma(\mathcal{D}_U)$	SA	ASR
$\mathcal{D}_U$	Clean	80.06 %	0.81 %
$\mathcal{D}_U$	0.1	72.22 %	100 %
$\mathcal{D}_L \bigcup \mathcal{D}_U$	Clean	81.25 %	99.98 %
$\mathcal{D}_L \bigcup \mathcal{D}_U$	0.1	76.45 %	99.98%

OPTML

SafeAI

A Pilot Study



This presents the opportunity for designing a more careful self-training scheme to prevent backdoor attacks !

Additional clean unlabeled data to erase the backdoor effects from model

However, naive pseudo-labeling of poisoned data can nullify this effect

model trained with self-training under ison ratio of labeled portion of CIFAR- re-trained on poisoned labeled portion 00 %)

	$\gamma(\mathcal{D}_U)$	SA	ASR
$\mathcal{D}_U$	Clean	80.06 %	0.81 %
$\mathcal{D}_U$	0.1	72.22 %	100 %
$\mathcal{D}_L \bigcup \mathcal{D}_U$	Clean	81.25 %	99.98 %
$\mathcal{D}_L \bigcup \mathcal{D}_U$	0.1	76.45 %	99.98%

OPTML

SafeAI

Alleviating Backdoor Via Data Augmentation

(Shen et al. 22)

- Data Augmentations make it **harder** for the model to overfit to **"easy to learn but bad" features**

(Borgnia et al. 21)

- In some cases, it has been shown to even **prevent backdoor in supervised settings**

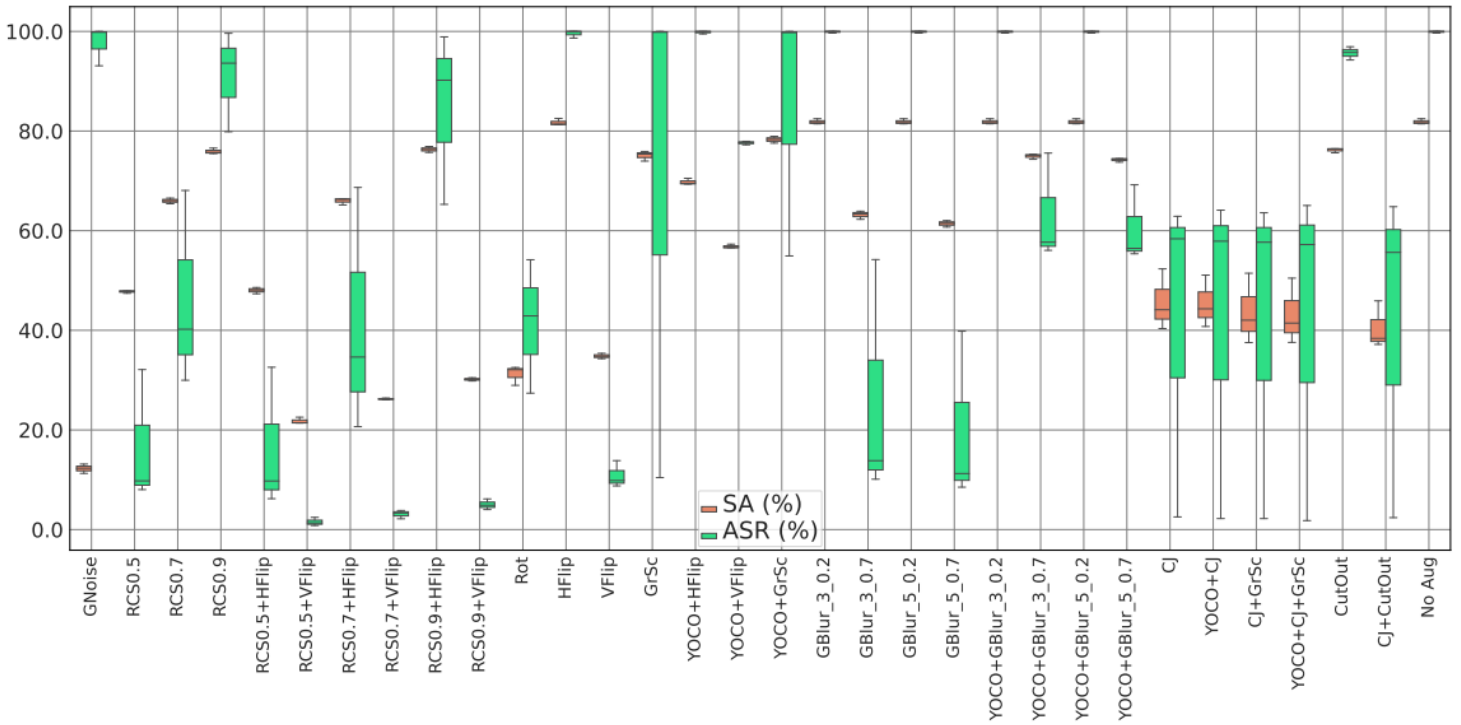
R. Shen, S. Bubeck, S. Gunasekar, Data augmentation as feature manipulation, in: Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 19773–19808

E. Borgnia, V. Cherepanova, L. Fowl, A. Ghiasi, J. Geiping, M. Goldblum, T. Goldstein, A. Gupta, Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 3855–3859.

OPTML

SafeAI

Alleviating Backdoor Via Data Augmentation

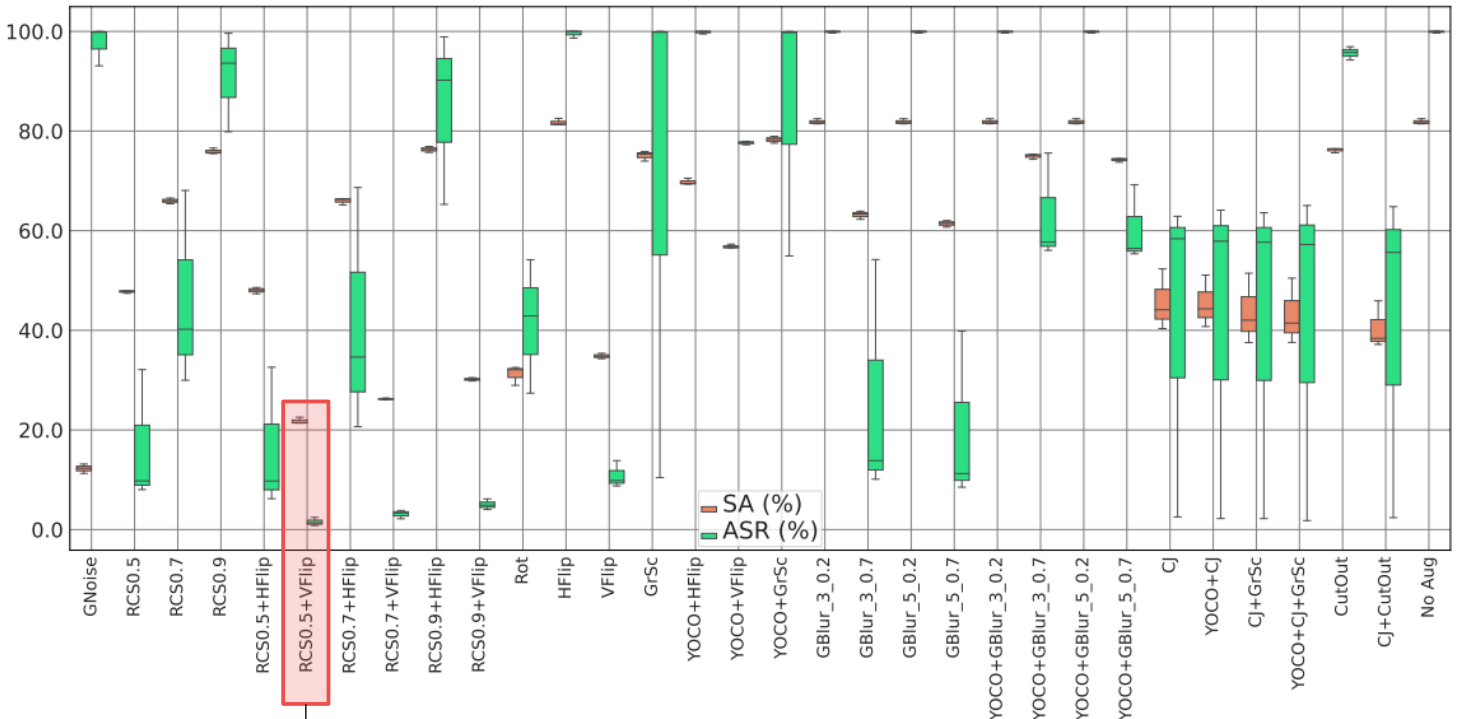


- Data Augmentations make it harder for the model to overfit to “easy to learn but bad” features
- In some cases, it has been shown to even prevent backdoor in supervised settings
- Boxplot over Badnet Grayscale attack, BadNet RGB, Clean Label attack

OPTML

SafeAI

Alleviating Backdoor Via Data Augmentation



Lowest ASR; chosen for future algorithms

- Data Augmentations make it harder for the model to overfit to "easy to learn but bad" features
- In some cases, it has been shown to even prevent backdoor in supervised settings
- Boxplot over Badnet Grayscale attack, BadNet RGB, Clean Label attack

OPTML

SafeAI

Self-training with data augmentations

Poison ratio : 1%/10%

Train model with labeled data $\mathcal{D}_L$

Pretrained / Backdoored model

Infer Pseudolabels on data

Retrain model with dataset

Make dataset using images + pseudolabel

Labeled dataset $\mathcal{D}_L$ transformed with data augmentation

Unlabeled dataset $\mathcal{D}_U$

Model

Pseudolabels

OPTML

SafeAI

Self-training with data augmentations

Trained on $\mathcal{D}_L$

Self-Training without data augmentations

Dataset	Backdoor Attack	$\gamma(\mathcal{D}_U)$	Pretrained Model		Semi-supervised Baseline		Proposed Method	
			SA	ASR	SA	ASR	SA	ASR
CIFAR-10	BadNet Gray-Scale	0.1	**81.65 %**	100 %	75.32 %	100 %	70.45 %	**75.02 %**
	BadNet RGB		**81.45 %**	100 %	76.45%	99.98 %	70.42 %	**50.90 %**
	BadNet Gray-Scale	0.01	**81.65 %**	100 %	80.85 %	100 %	73.37 %	**2.40 %**
	BadNet RGB		**81.45 %**	100 %	80.48 %	100 %	71.49 %	**4.98 %**
	Clean-Label Attack	0.25	**82.45 %**	99.63 %	81.40 %	98.57 %	73.39 %	**44.90 %**
CIFAR-10 + 500K TinyImages	BadNet Gray-Scale	0.01	**94.32 %**	100 %	89.09 %	99.98 %	84.81 %	**14.41 %**
	BadNet RGB		**94.70 %**	100 %	90.19 %	100 %	84.19 %	**14.33 %**
	Clean-Label Attack	0.05	**93.78 %**	99.07 %	89.43 %	91.19 %	83.67 %	**20.61 %**

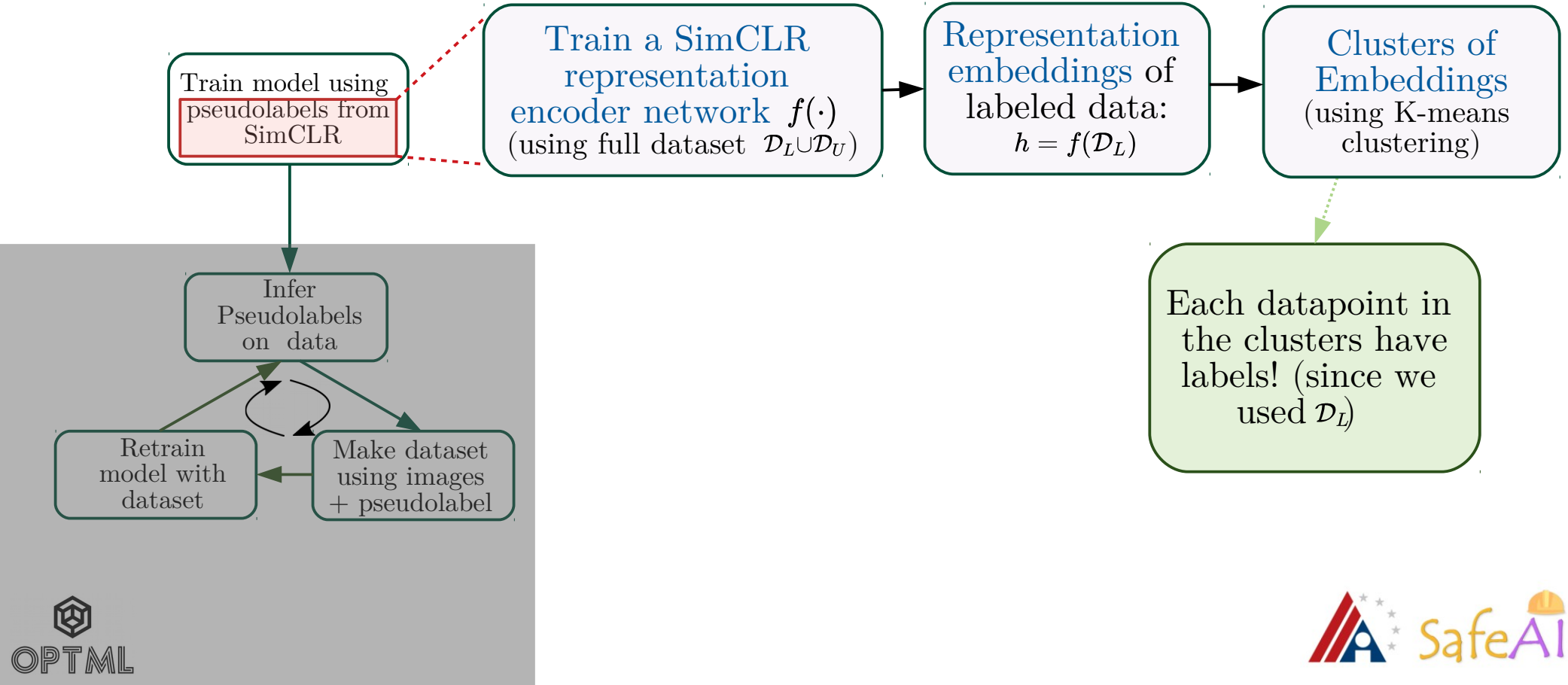
SA may be gained from $\mathcal{D}_U$ (even with strong data augmentation)

Successful in **combating backdoor**

OPTML

SafeAI

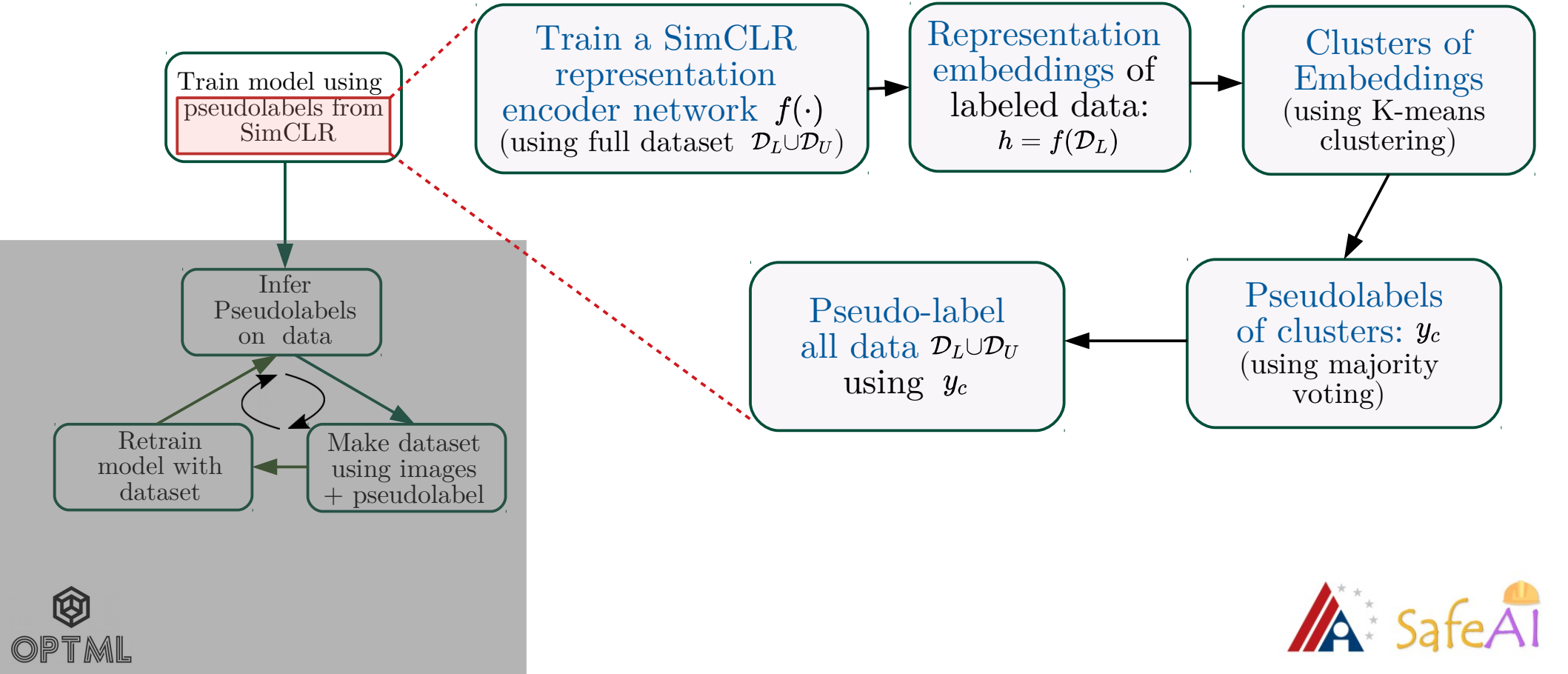
Self-training with SimCLR



OPTML

SafeAI

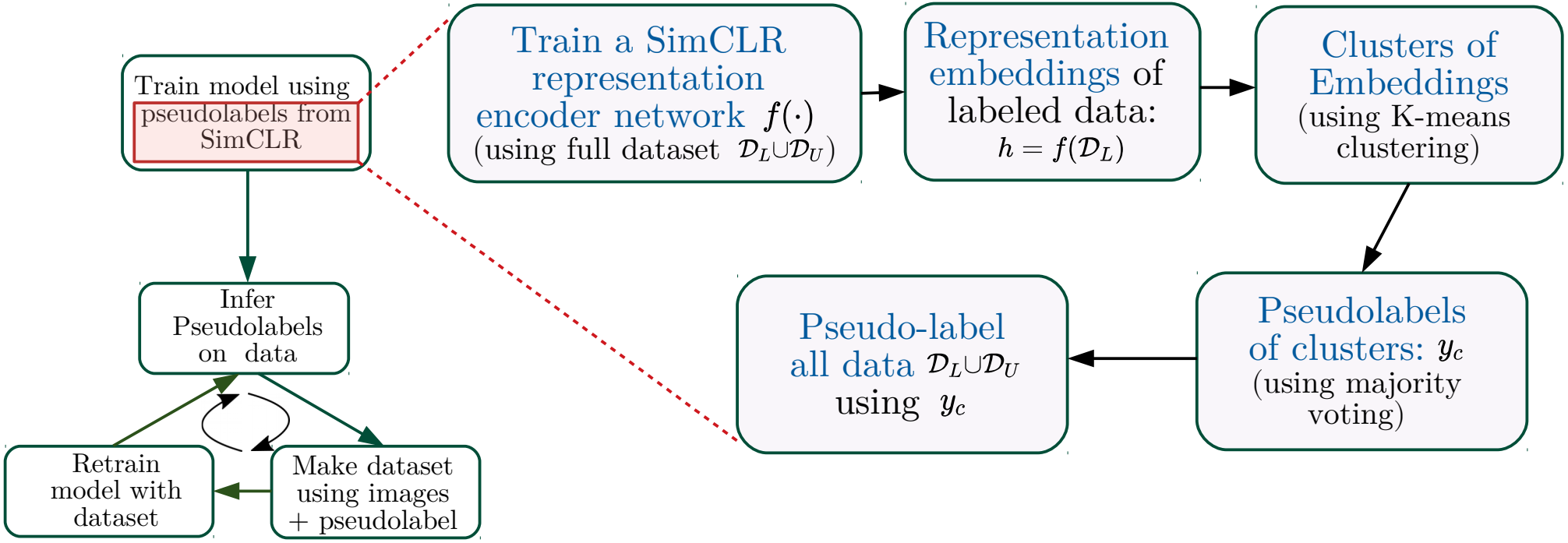
Self-training with SimCLR



OPTML

SafeAI

Self-training with SimCLR



OPTML

SafeAI

Self-training with SimCLR

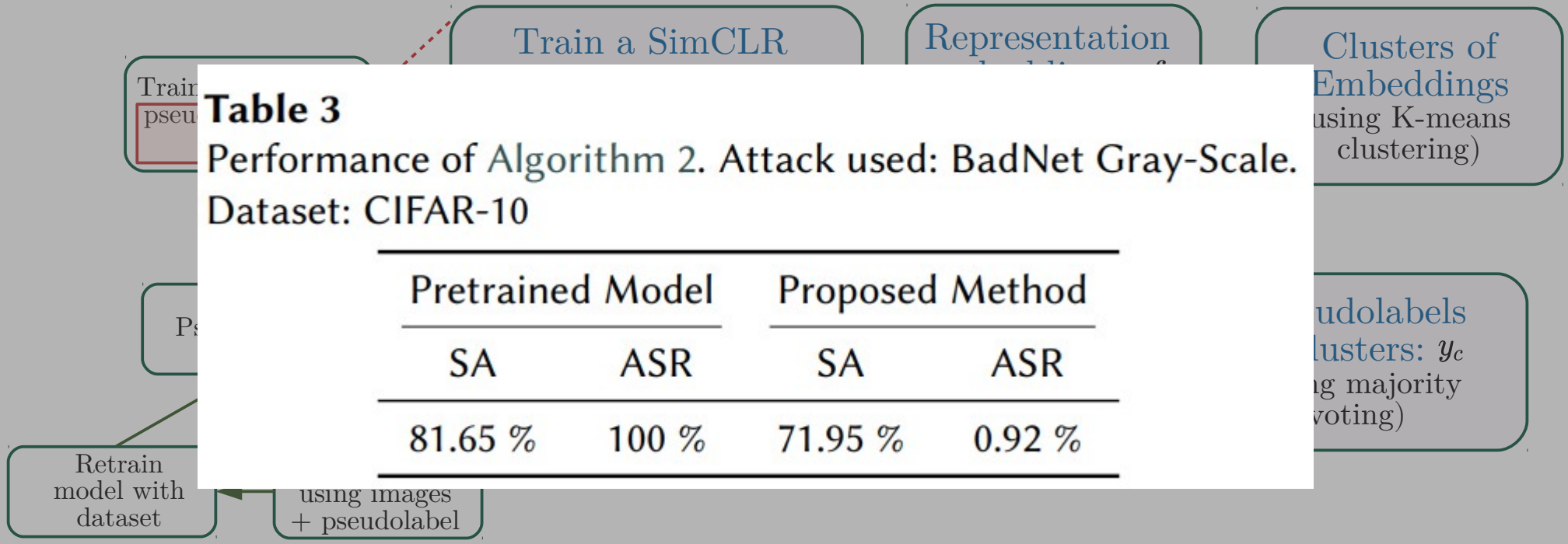


Table 3
Performance of Algorithm 2. Attack used: BadNet Gray-Scale. Dataset: CIFAR-10

Pretrained Model		Proposed Method	
SA	ASR	SA	ASR
81.65 %	100 %	71.95 %	0.92 %

OPTML

SafeAI

Conclusion

- ✔Step towards **understanding** the potential of **self-training** as a learning paradigm for **backdoor mitigation**

OPTML

SafeAI

Conclusion

- ✔Step towards **understanding** the potential of **self-training** as a learning paradigm for **backdoor mitigation**
- ✔Potential development of **trigger-agnostic** augmentation leveraging the **self-training framework**
  - to reduce ASR
  - to maintain SA

OPTML

SafeAI

Thank You!

Questions ?

Contact: Soumyadeep Pal

(soumyade@ualberta.ca)

OPTML

SafeAI