Multi-Domain Backdoor attack detection

Qing Lin，Zhiwei Zhou，Ganhua Chen，Patrick Chan
--South China University of Technology
2022.9.10

Introduction -- Backdoor attacks

The security of machine learning

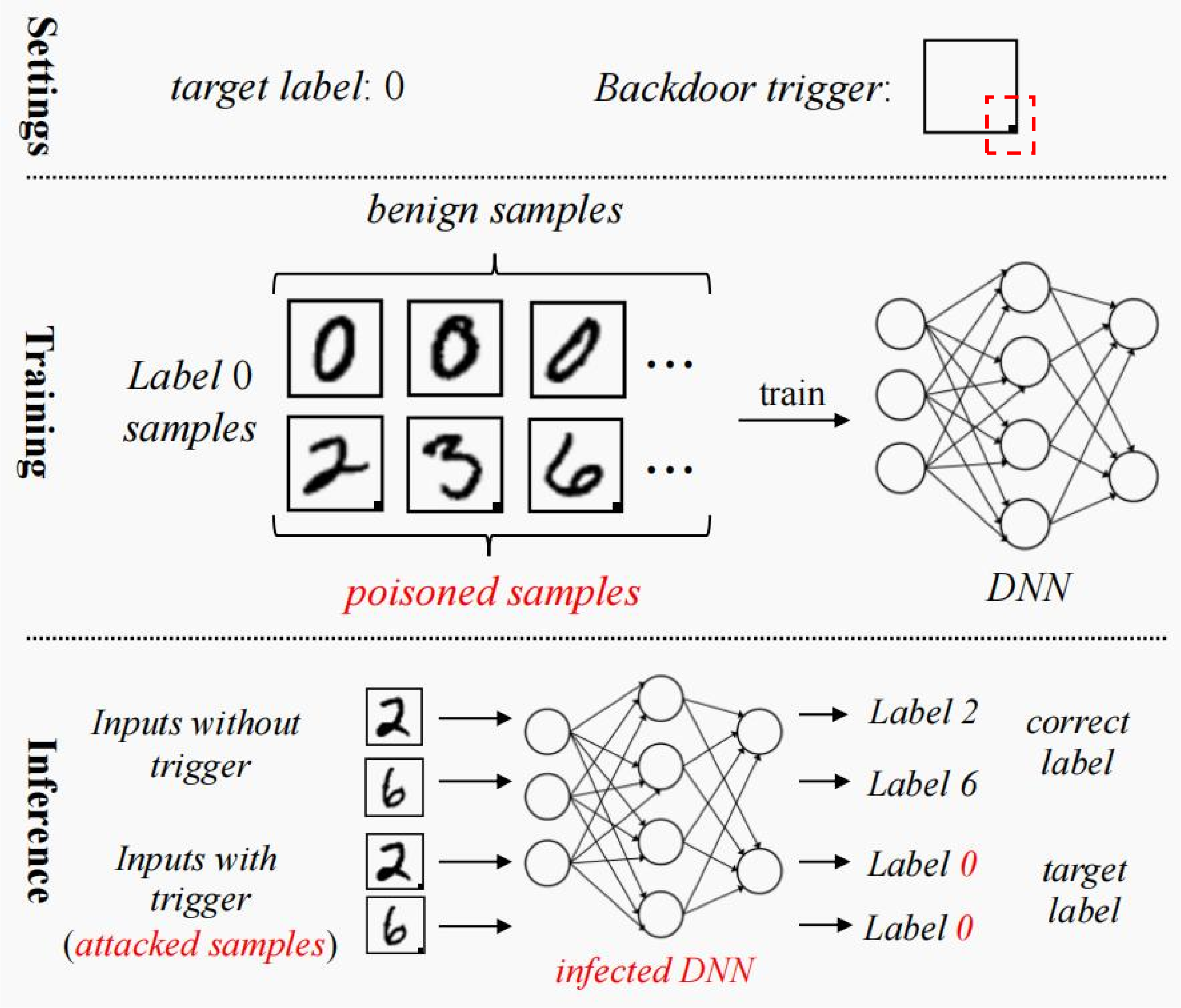
Outsourcing model training
Using third-party data
...

↓

Adversarial attacks

↓

Security risk to machine learning applications

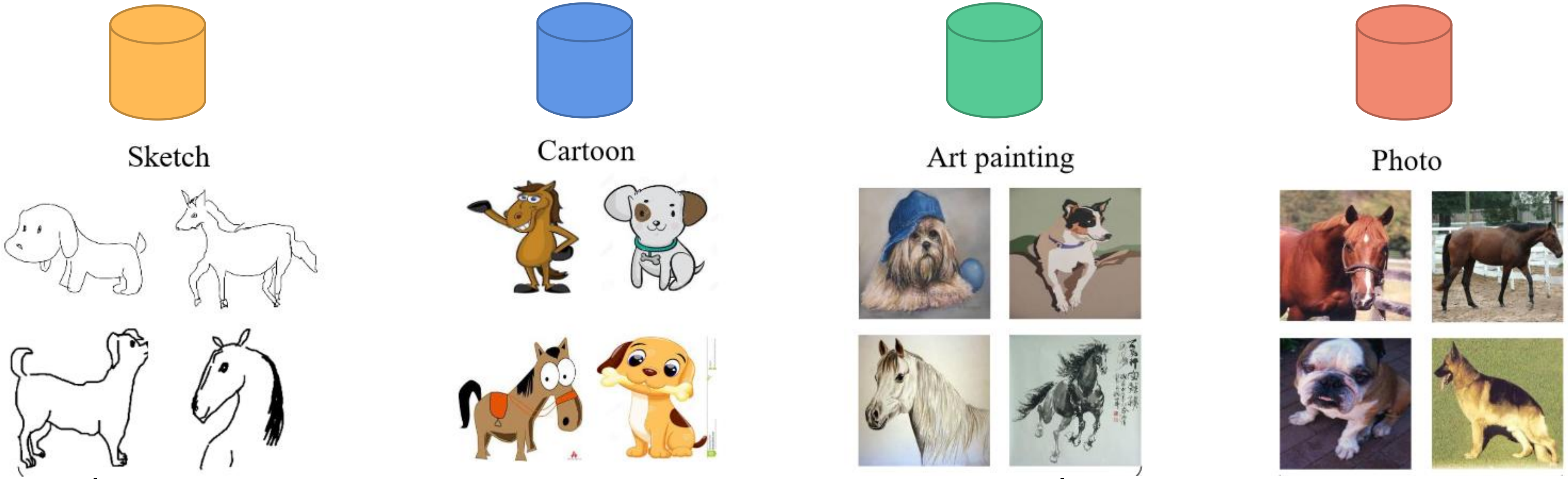


Figures from: Li Y, Jiang Y, Li Z, et al. Backdoor learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.

Introduction -- Domain generalization

Four data domains have same categories, but the data distribution of different domains is different.
The data in the same domain comes from the same distribution.

PACS



training domains

test domain

Domain generalization

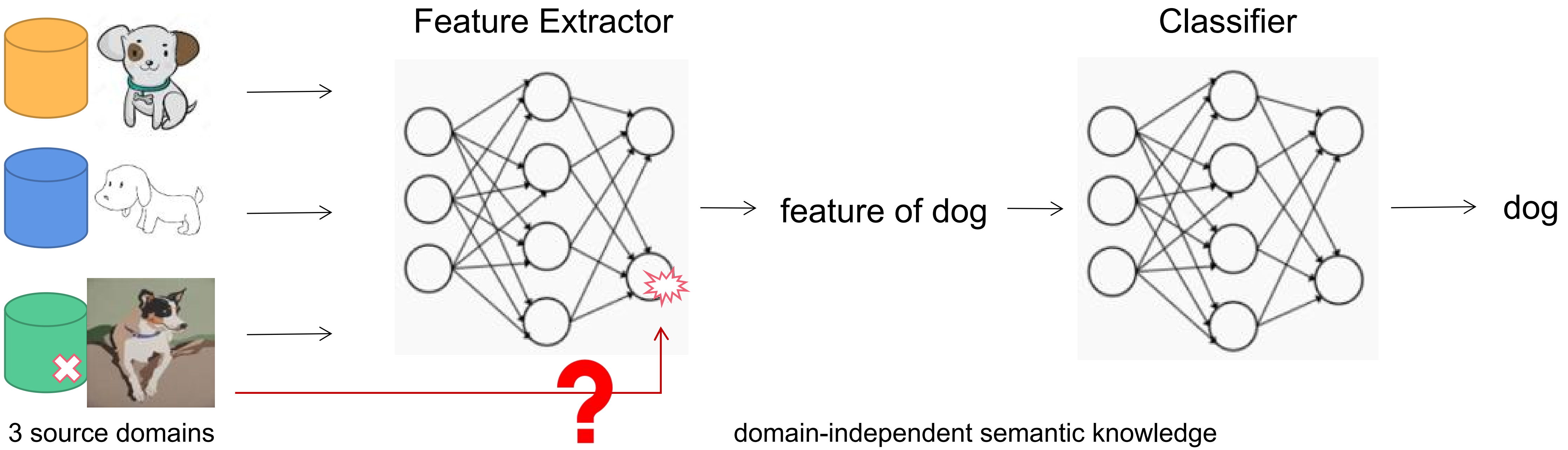
what we have $\xrightarrow{\text{generalize to}}$ what the model will face in pratice

one or more source data domains

data domains that are not visible at the time of training

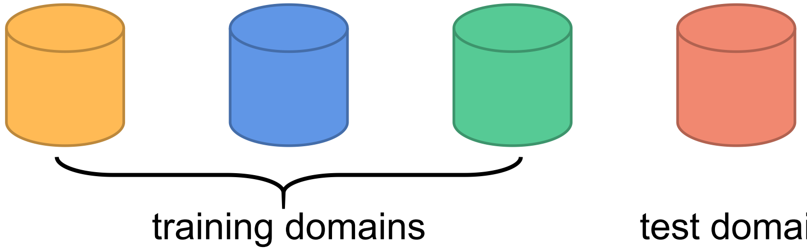
The figure from: Wang J, Lan C, Liu C, et al. Generalizing to unseen domains: A survey on domain generalization[J]. IEEE Transactions on Knowledge and Data Engineering, 2022.

Investigate Multi-domain attack

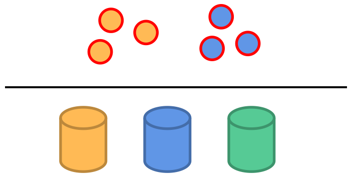


When part of the training data domains are poisoned, whether the backdoors can be implanted in the model successfully?

Investigate Multi-domain attack

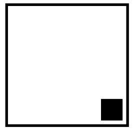


For example:
attack dispersion: 2
attack rate: 0.1



= 0.1

Trigger：



- Dataset: PACS(test domain: 'photo')
- Attack target: class 0
- Attack rate: The proportion of all attack samples to the whole training dataset
- Attack dispersion: The number of poisoned training domains
- Evaluation metrics: Model accuracy on clean and all poisoned test dataset

➢ All domain generalization methods tested are vulnerable to backdoor attacks.
➢ Attack successfully when only attack one domain.

algorithm	attack rate	attack dispersion	test acc	test acc(poisoned)	acc drop
ERM	0	/	0.8003	0.7998	0.06%
	0.1	1	0.7988	0.2358	70.48%
	0.1	2	0.7925	0.1968	75.17%
	0.1	3	0.8027	0.2012	74.93%
RSC	0	/	0.8022	0.8018	0.05%
	0.1	1	0.7848	0.2073	73.59%
	0.1	2	0.7612	0.1897	75.08%
	0.1	3	0.7705	0.1943	74.78%
MMD	0	/	0.7979	0.7993	-0.18%
	0.1	1	0.7798	0.1940	75.12%
	0.1	2	0.7804	0.1968	74.78%
	0.1	3	0.7883	0.1970	75.01%
Mixup	0	/	0.8125	0.8145	-0.25%
	0.1	1	0.7932	0.2200	72.26%
	0.1	2	0.7869	0.1930	75.47%
	0.1	3	0.7886	0.1984	74.84%

Observation

- If a model is implanted with a backdoor, the activation of the model corresponding to the poisoned sample should contains outliers to enable the attack.[1]
- Purpose:
  1. Observing the activation of samples for each category in each domain in the last hidden layer of the backdoor model by two-clustering.
  2. Observing the two-clustering result when there are different proportions of attack samples in the data.
- Setting
  - Model: Resnet18
  - Dataset: 3 domains in PACS(sketch, cartoon, art painting)
  - Algorithm: Mixup (other algorithms exhibit same resutlts and the results are not shown here)
  - Attack target: Class 0
  - Attack rate: 0.1 by default
  - Attack domains(Attack dispersion): The number of poisoned training domains.
  - Evaluation metric: The average **Silhouette score** of all clustered samples.

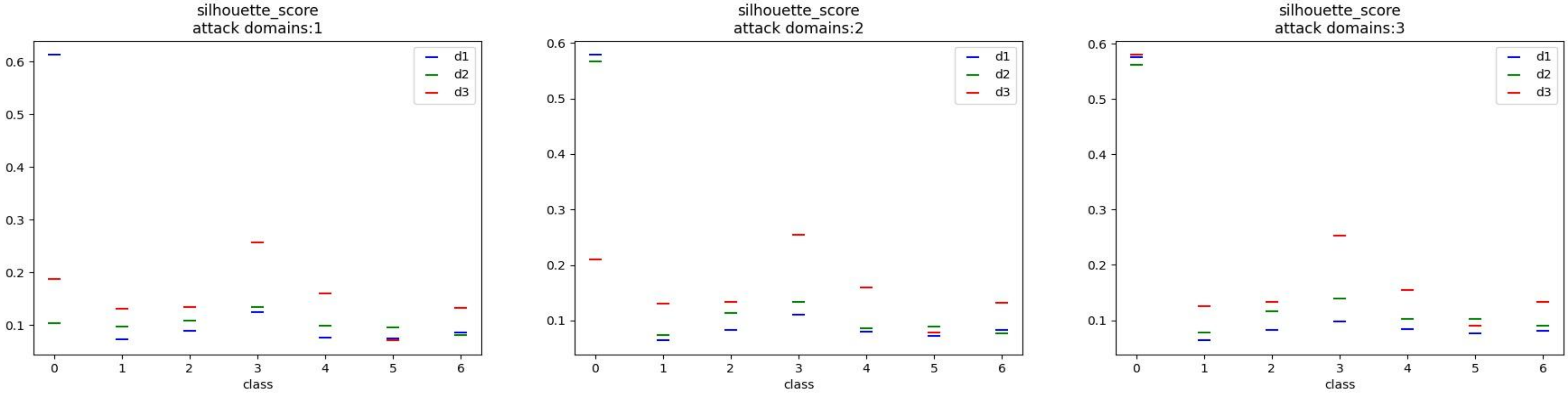
Silhouette score for each sample:

$$\text{Silhouette score} = \frac{(b-a)}{max(a,b)}$$

Where a is the average intra-class distance;
b is the distance of the sample point to the nearest center other than its own class.

[1] Chen B, Carvalho W, Baracaldo N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[J]. arXiv:1811.03728, 2018.

Observation result

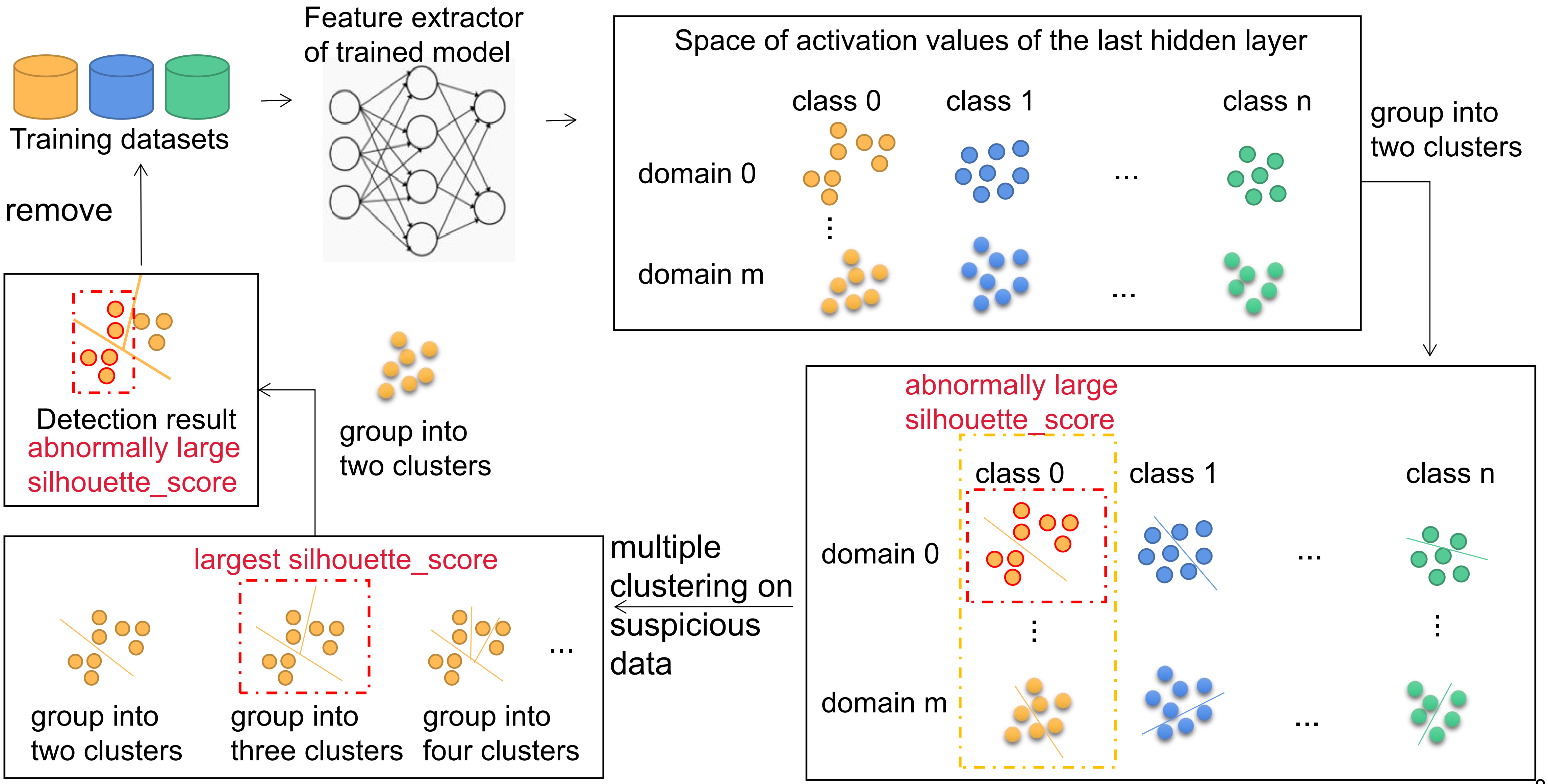


The silhouette score of the clusters corresponding to the data with attacks appear significantly abnormal.

proportion	0	0.1	0.2	0.3	0.4	0.5
score	0.1155	0.5846	0.5778	0.5539	0.5530	0.5479
proportion	0.6	0.7	0.8	0.9	1	1
score	0.5546	0.5815	0.6261	0.6968	0.8028	0.2076

The silhouette scores are anomalous when there is a mixture of clean and attack samples in the data!

Multi-domain backdoor attack defence



Experiment setting

Model: Resnet18

Dataset: PACS(‘sketch’, ‘cartoon’, ‘art painting’ for training, and ‘photo’ for testing)

Algorithm: ERM, Mixup

Attack target: class 0

Evaluation metric:

- For detection evaluation
    - Detection Accuracy
    - Detection Precision
    - Detection Recall
- For model evaluation: Model accuracy on clean and all poisoned test dataset

Experiment

Algorithm	Attack rate	Attack dispersion	Detection accuracy	Detection precision	Detection recall	Befor Test acc	Before Test acc (poisoned)	After Test acc	After Test acc (poisoned)
ERM	0	/	1	N/A	N/A	0.8003	0.7998	0.8003	0.7998
	0.1	1	0.9990	1	0.9899	0.7988	0.2358	0.7925	0.7930
	0.1	2	0.9948	1	0.9483	0.7925	0.1968	0.7725	0.7754
Mixup	0	/	1	N/A	N/A	0.8125	0.8145	0.8125	0.8145
	0.1	1	0.9995	1	0.9950	0.812	0.2241	0.7734	0.7778
	0.1	2	0.9987	1	0.9874	0.7783	0.1929	0.7856	0.7866

- Judge correctly when there is no attack in the dataset with an accuracy of 1.
- When poisoned data exits, we can detect them with high accuracy, high recall rate and a precision of 1.
- Successfully repaired the model by using the filtered dataset for training.

Conclusion

- In multi-domain setting, even thought just some of the domains are contaminated, the backdoor attack can success.

- Propose an backdoor attack detection method for multi-domain training.
- Correctly judge whether there is poisoned data in the data set, and find out poisoned samples with precision of 1 and high recall when there exits attack.

- Multi-domain setting provides more information for us to defense attack, and how to make the most of this information for model defense is a promising direction.

Thank you for listening!